PREDICTION OF RESERVOIR FLUID PROPERTIES USING MACHINE LEARNING

by

Munirudeen Ajadi Oloso

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth

JUNE 2018

Abstract

The phase and volumetric behaviour of reservoir fluid properties, referred to as pressure-volume-temperature (PVT) properties, involve the thermodynamic studies of the fluid with respect to pressure, temperature and its volumetric compositions. PVT properties are usually determined by laboratory experiments performed on the actual samples of the reservoir fluid. Failing that, these fluid properties have been evaluated by some other methods such as equations of state, empirical correlations and recently, machine learning models.

Machine learning is basically the prediction of the future with, (supervised learning), or without, (unsupervised learning), prior knowledge of the past. A common problem for the standalone machine learning technique is local minimum. In view of this, ensemble systems and hybrid techniques have been developed successfully for improvement in different fields.

This work introduces two different ensemble methods based on support vector regression and regression trees where both ensemble approaches utilise a novel concept tagged “Tying Ranking” in selection of the base models. Also, a hybrid system for reservoir fluid characterisation with a novel way of grouping petroleum fluid properties using intelligent method was developed. The hybrid system uses K-Means clustering for the intuitive grouping along with functional networks for the prediction.

The performance and generalisation of the developed models are compared against their standalone and selected empirical models using some statistical measures which are commonly used for performance evaluation in the petroleum industry.

In the first category of experimentation, the impact and effect of training the machine learning models with more diverse and bigger data set is shown. Effects of using different functional forms to predict dead oil, saturated and undersaturated viscosity are also explored. In addition, impacts of different statistical measures on the predicted outputs and wrong interpretations of results in the literature are examined.

The main statistical measures that are used for comparison are root mean squared errors, average absolute percentage relative error and maximum absolute percentage relative error. For each of the reservoir fluid properties considered in this work, at least one or more of the developed machine learning models have better overall and average performance than all the compared correlations in each category.

The superiority of the three developed machine learning models is visible in the trend analysis as they show less deviations in results compared to the empirical correlations and their standalone methods in most cases for all the considered reservoir fluid properties.

Acknowledgement

"*Whoever is not grateful to people cannot be grateful to Allah*" (Abu Dawūd).

All praise is due to Allah for His abundant blessings on me, my parents and the entire family. Even if I had a billion tongues, they would not be sufficient to appreciate His favours and blessings.

I would like to thank the Petroleum Technology Development Fund (PTDF), Nigeria, for sponsoring my PhD study at the University of Portsmouth. I would also like to thank my first supervisor: Dr Mohamed Hassan and other supervisory team members: Dr Mohamed Bader and Dr James Buick, for their advice, supports, encouragement and guidance throughout this study. I am also thankful to Dr Jebraeel Gholinezhad and Dr Lateef Akanji, who acted as the internal and external examiners respectively for this work, for their constructive recommendations.

Special thanks and honour to my parents, Sheik Sirajudeen and Alhaja Dhikrah Oloso, for their encouragement and supports during this study. The upbringing they have given me and my siblings is exceptional. May Allah reward them abundantly in this world and the hereafter. Also, my heartfelt appreciation and gratitude to my dearest better half, Qudrah Adeola (Ummu Mabrūr) for her unwavering supports, and my lovely kids: Mabrūr, Hawwa, Abdul-Mannān ("*baba jęję of Leeds"*) and Thabit ("*baba jęję of Lagun"*) who occasionally missed their dad during this study,….daddy will not be sneaking out of the house to "engineering building" again. May Allah shower His mercies and blessings on you all - amin.

To all my elder and younger siblings, friends and colleagues who are too numerous to mention, I say "*Jazakumu-lLahu khayran*" to all of you. May Allah reward you for your continuous supports.

Declaration of Authorship

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Word Count: 32424

Table of Contents

List of Tables

List of Figures

Abbreviations

AAD: average absolute deviation

AAPRE: absolute average percent relative error

ANN: artificial neural network

ANFIS: adaptive neuro-fuzzy inference system

APE: average percentage error

CI: Computational Intelligence

CMS: constant mass study

CVDS: constant volume depletion study

FL: fuzzy logic

FN: functional network

FVF: formation volume factor

GA: genetic algorithm

GOR: Gas-oil ratio

MLP: multi-layer perceptron

NF: neuro fuzzy

PT: pressure-temperature

PLCIS: power-law committee with intelligent systems

PVT: pressure-volume-temperature

RBF: radial basis functions

RMSE: root mean squared error

SVM: support vector machine

SVR: support vector regression

Nomenclature

B formation volume factor, bbl/STB

B_o oil formation volume factor, bbl/STB

B_{ob} saturated/bubble point oil formation volume factor, bbl/STB

B_t total oil formation volume factor, bbl/STB

C coefficient of isothermal compressibility factor, psi^{-1}

C_o oil coefficient of isothermal compressibility factor, psi^{-1}

C_{ob} saturated/bubblepoint oil coefficient of isothermal compressibility factor, psi^{-1}

CC correlation coefficient

E_a average absolute percentage error

E_{max} maximum average absolute percentage error

P pressure

R ideal gas constant

RMSE root mean square error

P_b bubblepoint pressure

R_s solution gas/oil ratio in scf/STB

SD standard deviation

T temperature

V volume

γ_{API} oil gravity API

γ_g dissolved gas relative density (air=1)

γ_o stock tank oil relative density (water=1)

μ_o oil viscosity in centipoise (cp)

μ_{ob} oil viscosity in centipoise (cp)

μ_{oa} undersaturated oil viscosity in centipoise (cp)

Disseminations

Oloso, M.A., Hassan M.G., Buick J., Bader-El-Den M. (2016). Oil PVT characterisation using ensemble systems. In: Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. IEEE, pp 61–68

Oloso, M., Bader-El-Den, M., Buick, J., & Sayed, M. H. (2017, January). Hybrid functional networks for PVT characterisation. In *2017 Intelligent Systems Conference: IntelliSys 2017*. IEEE.

Oloso, M. A., Hassan, M. G., Bader-El-Den, M. B., & Buick, J. M. (2017). Hybrid Functional Networks for Oil Reservoir PVT Characterisation. *Expert Systems with Applications*.

Oloso, M. A., Hassan, M. G., Bader-El-Den, M. B., & Buick, J. M. (2017). Ensemble SVM for characterisation of crude oil viscosity. *Journal of Petroleum Exploration and Production Technology*, 1–16. http://doi.org/10.1007/s13202-017-0355-x

Chapter 1. Introduction

1.1. Background

Volumetric determination of reservoir fluid properties are very important in reservoir engineering. They are used for different computations such as material balance calculations, well test analysis, reserve estimates, and numerical reservoir simulations (Osman, Abdel-Wahhab, & Al-Marhoun, 2001). Ideally, these properties are determined by taking reservoir fluid samples in a controlled process for laboratory analysis. The laboratory process is marked with some notable shortcomings, prompting some alternative methods to be sought.

Some of the problems with the conventional laboratory experimentation include high set up cost; requirement for high level of scientific expertise and long turnover testing period (Malallah, Gharbi, & Algharaib, 2006). These problems associated with the laboratory experimentation have prompted the industry to seek alternative methods to obtain estimations of petroleum reservoir fluid properties which include, but not limited to, bubble-point pressure, oil formation volume factor, gas solubility, oil viscosity, specific gravity, and gas-specific gravity.

In addition to the laboratory experimentation, the two most commonly used methods of calculating pressure-volume-temperature (PVT) properties are equation-of-state (EOS) and empirical correlations. The EOS is considered computationally complex and requires extensive detailed compositions of reservoir fluids. On the other hand, PVT correlations do not usually require compositional elements, rather, they only involve simple mathematical computations. In short, they only require readily available reservoir data such as reservoir pressure, temperature, oil-gravity and gas-specific gravity. Several correlation methods have been reported in publications to estimate PVT properties such as (Al-Marhoun, 1988; Glasø, 1980; Lasater, 1958; Standing, 1947; Vazquez & Beggs, 1980). One commonly emphasised shortcoming of the empirical correlations is their sensitivity to new data sets since many of them have been developed using regional PVT data.

In recent years, some machine learning (ML)/artificial intelligence (AI) methods such as fuzzy inference systems, artificial neural networks (ANNs) and different evolutionary algorithms have been used when solving the petroleum engineering and geology problems. The petroleum engineering problems are highly complex and non-linear. The artificial neural network (ANN) model has been one of the attractive tools used in geo-engineering applications due to its satisfactory performance in the modeling of non-linear multivariate problems. The attractiveness of ANN comes from the information processing characteristics of the system, such as non-linearity, high parallelism, robustness, fault and failure tolerance, learning capability, ability to handle imprecise

and fuzzy information, and their capability to generalise (Jang, 1993). ANN-based models can provide practically accurate solutions for precisely or imprecisely formulated problems and for phenomena that are only understood through experimental data and field observations (Basheer & Hajmeer, 2000).

With a bid to improve the prediction of PVT properties, some ANN architectures have been proposed (Farshad, Garber, & Lorde, 2000; Gharbi, Elsharkawy, & Karkoub, 1999; Nguyen & Chan, 2005). What seems to be common among the researchers as the reasons for using ANN are its advantages which include its computational efficiency, non-linear characteristics, generalisation properties and ease of use with high-dimensional data. However, there is no doubt that ANN has its shortcoming in theses realms as getting a very stable and high performance ANN model is not always trivial. In this regard, fuzzy logic systems, Functional Networks (FN) and Support Vector Machines (SVM) among others have been strongly advocated to outperform ANN especially in terms of handling non-linear characteristics and uncertainties (El-Sebakhy et al., 2007; Oloso et al., 2009); Olatunji et al., 2010).

Different proposed optimisation algorithms that have been proved to outperform ANN also have the potential problem of local minimal where the chosen and final model may not necessarily be the most optimal one. This research is aimed to improve the prediction of some PVT properties by reducing the problem of potential local minimal possibly associated with some ML techniques.

In view of this, two ensemble ML models and a hybrid system have been developed in this study to predict some PVT properties. To date, no ensemble model has been utilised in modelling PVT properties despite its successful implementation in other fields. As there is no defined taxonomy in building an ensemble algorithm, a heuristic approach based on the error statistics that are usually utilised in analysing PVT properties has been used in developing the ensemble models. Also, some researchers have used API classification to develop different correlations for each group of the petroleum fluids. However, instead of the manual grouping, this thesis uses an intelligent system to group the entire input fluids before passing the clusters into a ML algorithm for prediction.

1.2. Methods for Determination of PVT Properties

The existing methods that are used to determine PVT properties can be grouped into four: laboratory experimentations; using equations-of-state; empirical correlations and ML/computational intelligence (CI) techniques.

- Laboratory experimentations: All other methods partially or entirely depend on this at least to get the sample data for model simulation or equation tuning or regression. It is initiated

by sampling of the reservoir fluids dictated by the intended experiment to be carried out or the experiments to be performed. Typical sampling methods include: bottom hole sampling; surface recombination sampling and repeat formation tester sampling. Typical experimental procedures include: constant composition expansion; differential liberation experiment; multi-stage separator test and constant volume depletion experiment.

- Equations of state: The aim of different equations of state is to represent the thermodynamic behaviour of fluids mathematically with respect to pressure, temperature and volume. Most of the EOS that are used for characterising reservoir fluids are derived from van der Waal's equation (Van-der-Waals, 1873). Common EOS that are in reservoir fluid characterisation in this respect include: Redlich & Kwong (1949); Soave-Redlich-Kwong equation (Soave, 1972) and Peng & Robinson (1976).
- Empirical correlations: These usually involve modelling the relationship between a desired PVT property and some independent variables using multiple linear regression analysis. More details are provided in Chapter on this.
- ML/CI techniques: Some intelligent techniques have been implemented for prediction of different PVT properties. Chapter 3 discusses and review some of these techniques with respect to prediction of PVT properties.

1.3. Aims of the Thesis

The aim of this research is to test and compare reliability, consistency and generalisation of selected empirical models against some standard ML techniques, newly developed ensemble and hybrid models. Thorough comparison of results is done using graphical and tabular representations.

1.4. Objectives of the Thesis

To fulfil the aims of this work, it is necessary to achieve the following objectives:

1. To build ensemble models based on regression tree (RT) and support vector regression (SVR) to predict some PVT properties.
2. To develop a hybrid system using K-Means clustering and functional networks (FN) for prediction of some PVT properties.
3. To compare the results of the ensemble models and the hybrid system with their standalone techniques and some common empirical models.
4. To investigate reliability, prediction accuracy and generalisation of all the developed ML techniques and some empirical correlations on the available data sets.
5. Limitations of the developed models especially with respect to the available and new data sets used for evaluation are observed.

1.5. Thesis Outline

This chapter discusses the background and motivation for embarking on this research work. It highlights the aims and objectives which have been pursued in the research work.

In chapter 2, a review of the empirical correlations has been done. A tabular presentation of the review has been done with emphasis on the API range, regional source of the data set, and the reported evaluation error reported by the authors in the reviewed literature.

Chapter 3 discuses theoretical backgrounds of some relevant machine learning/computational intelligence techniques. This is then followed by a review of some of these techniques or their derivatives which have been applied to the prediction of PVT properties.

Chapter 4 develops two new different ensemble techniques based on SVR and regression trees (RTs). Also, a hybrid system of K-Means clustering and functional network (FN), inspired by the API grouping of petroleum fluids, was introduced.

In chapter 5, results and discussions on the performances of the hybrid functional network, the ensemble systems, their standalone ML techniques and some empirical correlations have been discussed. Explanations are given on the impacts of the statistical measures used in this study and their wrong usage in the literature.

Chapter 6 summarises the contributions of this research work. A concise appraisal of the evaluated methods has also been done. Possible directions for future work and improvement have also been given.

Chapter 2. Empirical Correlations for Prediction of Reservoir Fluid Properties

2.1. Overview of Empirical Correlations for Reservoir Fluid Characterisation

In this section, important black oil properties are explored. Common correlations that are used to predict these properties are also reviewed. The process of developing the correlations usually involve linear or nonlinear regression analyses to obtain equation(s) that best fit the training data and minimise the deviation from the actual measurements. In other words, if Y represents the target data and $\hat{Y}$ is the predicted output, then the aim is to obtain:

$$\min(Y - \hat{Y}) \tag{2.1}$$

A plethora of correlations have been developed for different PVT properties, especially, bubblepoint pressure (P_b) and oil formation volume factor (oil FVF/B_o). Some authors have developed more than one correlation for a given PVT property based on API (γ_{API}) grouping, usually, for $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$ (Vazquez & Beggs, 1980).A concise review of the common and recent correlations is presented. The PVT properties that will be considered are: P_b, B_o, viscosity (μ_o) and oil compressibility factor (C_o).

2.1.1. Correlations for Bubblepoint Pressure

A concise review of some empirical correlations for the prediction of bubblepoint pressure is presented. Most of the selected correlations in this review do not use oil compositions as inputs. Many of them use the input variables proposed by Standing (1947) which is based on the following functional form.

$$P_b = f(R_s, \gamma_g, \gamma_{API}, T) \tag{2.2}$$

The other common functional form is:

$$P_b = f(R_s, \gamma_g, \gamma_o, T) \tag{2.3}$$

Table 2.1. Review of Bubblepoint Pressure Empirical Correlations

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Standing (1947)	16.5-63.8	North America (California)	105	105 experimentally determined P_b values from 22 different crude oil mixtures were used for the study. An arithmetic average error of 4.8% was reported.

Lasater (1958)	17.9-51.1	North and South America	158	The correlation was based on 158 experimentally measured bubblepoint pressures of 137 independent systems. Average error of 3.8% was reported for the data.
Glasø (1980)	23.7-45.2	North Sea, Middle East, Algeria and United States	45	Corrections for the effect of non-hydrocarbon component in the crude oil were incorporated in the correlation. It was stated that the presence of the following compositions affect saturation pressure in different ways: CO_2, N_2 and H_2S. Very high CO_2 will make a correlation to underestimate P_b. For the effect of N_2 gas, high P_b will be required to force it into solution in the liquid phase of the hydrocarbon. Secondly, the *P-T* phase diagram for a system with N_2 has a lower envelope than a system without N_2. For the H_2S, only the γ_{API} of the stock tank oil changes its effect on the saturation pressure.
Vazquez and Beggs (1980)	N/A	Worldwide	5008	This correlation is based on the reported R_s correlation.
Al-Marhoun (1988)	19.40-44.6	Middle East	160	PVT analyses of 69 different Middle East oil/gas mixtures were used for this study. R^2 for the developed correlations is 0.997 with AAPRE of 3.66%. This was compared with some previous correlations, Standing (1947) [$R^2 = 0.979,\ APPRE = 12.08\%$] and Glasø (1980) [$R^2 = 0.891, APPRE = 25.22\%$].
Kartoatmodjo and Schmidt, (1991)	14.4-59.0	Indonesia, North America, Middle east and Latin America	5392	The correlation has two different set of coefficients for $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$. Performance of the correlation was compared with some published correlations.

Dokla and Osman (1992)	28.21-40.3	UAE	51	They reported AAPRE of 7.610% and claimed that there was no universal correlation as the properties of crude oil vary from region to region. The developed correlation was reported to perform better than all the compared correlations (Standing, 1947; Glasø, 1980; Al-Marhoun, 1988)
Omar and Todd (1993)	26.6—53.2	Malaysia	93	Standing correlation was used as the basis for the new model, regressing the equation based on the Malaysian crude oil to re-calculate the coefficients of the equation. AAPRE = 7.17% and correlation coefficient = 0.95.
De Ghetto et al. (1995)	6-22.3	-	1200	The correlation was developed based on Standing's correlation[7] for heavy and extra heavy crude oils with APPRE of 9.1% and 10.2% for API<10 and $10 < API \leq 22.3$ respectively.
Almehaideb (1997)	30.9-48.6	UAE	62	APPRE = 4.997% while other compared correlations gave higher APPRE ranging between 15.71% and 22.44%. Correlation coefficient = 0.989.
Hanafy et al. (1997)	17.8-47.7	Egypt	324	The initial solution gas/oil ratio is used as part of the independent correlating variable instead of R_{sb} as done by many other researchers. AAPRE = 18.59% and correlations coefficient = 0.9
Petrosky Jr and Farshad (1998)	16.3-45.0	Gulf of Mexico	90	AAPRE of 3.28% was reported for the correlation.
McCain Jr et al. (1988)	12-55	N/A	728 (Additional 547 data set for testing)	A non-parametric regression model was developed. AAPRE of around 13% was reported for the testing data set. Many other correlations and methods were compared and it was noted that none of them has lower

				error. The authors emphasised that APPRE of 13% could result in a 25% error in P_b prediction which is unacceptable, hence, the recourse to the traditional way of estimating P_b using laboratory measurement.
Khairy et al. (1998)	30.7-54.3	Egypt	43	AAPRE = 14.15% which was better than the results of all the compared eight previous correlation equations.
Velarde et al. (1999)	12-55	N/A	728	They derived their correlations based on the "reduced" PVT properties. AAPRE = 11.5%.
Al-Shammasi (2001)	6.0-63.7	World wide	1243	Linear regression was used to develop the new correlation. AAPRE of 17.85% and correlation coefficient of 0.9987 were reported in testing the global data. The author re-calculated the coefficients of many of the existing correlations to improve their performance in estimating the global PVT data. According to the author, all the compared correlations gave lower performance than the new correlation.
Boukadi et al. (2002)	29.5–45.0	Middle East	32	24 data set was used to generate the correlation while the remaining were used for testing it. AAPRE = 15.0460%. (Standing, 1947) gave a better result with APPRE of 12.1180.
Elsharkawy (2003)	N/A	Middle East	60	The derived model was compared with SRK-EOS and PR-EOS. For the actual data used for developing the model, AAD of 6.36, 10.91 and 9.23 were reported for the new model, SRK-EOS and PR-EOS respectively. However, SRK-EOS in some cases gives better performance than the model when the three models were tested on some external data according to the report.

Valko and McCain (2003)	6.0-63.7	World wide	1745	The authors used the GRACE method to develop a new correlation for the P_b. The authors suggested that it was not necessary to perform correlations based on a specific geographical region. Rather, a carefully well prepared universal correlation gives adequate results. AAPRE = 10.9%
Dindoruk and Christman (2004)	14.7-40	Gulf of Mexico	104	AAPRE = 5.70%
Malallah et al. (2006)	14.3-59	World wide	5200(another 234 data set was used for testing)	The authors used a non-parametric optimisation method called alternating conditional expectation (ACE) to develop the correlation. In testing, AAPRE of 17.31% and correlation coefficient of 0.9571 were reported.
Hemmati and Kharrat (2007)	18.8-48.34	Iran	287	AAPRE = 3.71% and correlation coefficient = 0.993
Khamehchi et al. (2009)	33.4-124	Unknown	94	Correlation coefficient = 0.93
Arabloo et al. (2014)	6.00-56.80	World wide	750+	AAPRE = 18.9% and correlation coefficient = 0.86. They compared their model with fifteen other correlations and none of them gave better result with respect to the two previously stated criteria. A group analysis was also performed where γ_{API} is used to divide the data sets into seven groups and analysed the performance and consistency of the evaluated correlations. The authors reported that their new correlation model gave the best consistency in prediction in all the seven groups.
Jarrahian et al. (2015)	14.08-51.79	Iran	207	Two different correlations were developed: compositional and non-compositional models. The non-compositional model was

				tested on another 1840 data sets from the literature with AAPRE of 15.555%. The compositional correlation was tested against 171 data set from the literature with AAPRE = 7.7%.

2.1.2. Correlations for Oil Formation Volume Factors

A concise review of some empirical correlations for the prediction of oil FVF at bubblepoint pressure is presented in Table 2.2. There are correlations for bubblepoint oil FVF and total FVF (B_t). Generally, the prediction of bubblepoint oil FVF is more accurate than the predictions of P_b and total oil FVF with respect to the evaluation statistical measures (Al-Marhoun, 2004). Two common functional forms that are used in B_{ob} regression analysis are as follow.

$$B_{ob} = f(R_s, \gamma_g, \gamma_{API}, T) \quad (2.4)$$

$$B_{ob} = f(R_s, \gamma_g, \gamma_o, T) \quad (2.5)$$

For the total oil FVF, two common functional forms are:

$$B_t = f(R_s, P, \gamma_g, \gamma_o, T) \quad (2.6)$$

$$B_t = f(P_b, P, \gamma_g, \gamma_o, T) \quad (2.7)$$

Table 2.2. Review of Empirical Correlations for Oil FVF

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Standing (1947)	16.5-63.8	North America (California)	105	Graphical correlation was developed using 22 different crude oil mixtures for B_o at P_b. An arithmetic average error of 1.17% was reported.
Vazquez and Beggs (1980)	N/A	Worldwide	6000	A correlation was developed for B_o at and below P_b. A correlation with two different set of coefficients was developed because of the wide variations in the volatility of the crude samples. The two groups were for $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$.

Glasø (1980)	23.7-45.2	North Sea, Middle East, Algeria and United States	45	Corrections for the effect of non-hydrocarbon component in the crude oil were incorporated in the correlation.
Obomanu and Okpobiri (1987)	14.98-43(γ_o:0.811-0.966)	Nigeria	503, 110	Two different correlations for different ranges of γ_o. 503 data points were used to develop a correlations for $0.811 \leq \gamma_o < 0.876$ while 110 data points were used to develop another correlation for $0.876 \leq \gamma_o \leq 0.966$. The reported AAPREs are 2.178% and 1.178% respectively, and the corresponding correlation coefficients in testing the models are 0.973 and 0.907 respectively.
Al-Marhoun (1988)	19.40-44.6	Middle East	160	PVT analyses of 69 different Middle East oil/gas mixtures were used for this study. R^2 for the developed correlations is 0.997 with AAPRE of 3.66%. This was compared with some previous correlations, Standing (1947) [$R^2 = 0.979,\ APPRE = 12.08\%$] and Glasø (1980) [$R^2 = 0.891, APPRE = 25.22\%$].
Kartoatmodjo and Schmidt, (1991)	14.4-59.0	Indonesia, North America, Middle east and Latin America	5392	Similar error and sensitivity analysis that was performed for the P_b prediction was carried out.
Dokla and Osman (1992)	28.21-40.3	UAE	51	AAPRE =1.225%. The developed correlation was reported to perform better than all the compared correlations (Standing, 1947; Glasø, 1980 and Al-Marhoun, 1988).
Al-Marhoun (1992)	9.5-55.9(for B_{ob})	Worldwide (mainly from Middle East and North America)	4012/3711/4005	4012, 3711 and 4005 data points were used to develop three different correlations for B_o at, above and below bubblepoint pressure. AAPRE of 0.57, 0.28

	10.4-49.2 (for B_o above and below bubblepoint)			and 1.68 were reported for the B_o correlations at, above and below bubblepoint respectively, and the reported correlations coefficients are 0.9987, 0.9997 and 0.9995 respectively.
Omar and Todd (1993)	26.6—53.2	Malaysia	93	Similar to their P_b model, Standing correlation was used as the basis for the new model. AAPRE = 1.44% and correlation coefficient = 0.99.
Almehaideb (1997)	30.9-48.6	UAE	62	AAPRE = 1.35%. According to the author, other compared correlations gave higher AAPRE, ranging between 1.87% and 4.91%. Correlation coefficient = 0.9985
Hanafy et al. (1997)	17.8-47.7	Egypt	324	The B_{ob} was correlated as a function of the initial solution gas/oil ration, rather than R_{sb}. AAPRE = 3.56% and correlation coefficient = 0.95.
Elsharkawy and Alikhan (1997)	19.9-42.76	Kuwait	171	AAPRE = 1.43% and correlation coefficient=0.9991. This model was reported to have better performance than all the compared 11 previous correlations in predicting B_{ob} for Kuwaiti crude oil.
Petrosky Jr and Farshad (1998)	16.3-45.0	Gulf of Mexico	90	AAPRE = 0.64%
Khairy et al. (1998)	30.7-54.3	Egypt	43	A correlation equation was proposed for the B_{ob} estimation with AAPRE of 3.27% which has poorer performance than 4 of the 8 compared correlations. The authors in the same article proposed another correlation for B_{ob} based on the weight average of the correlation results from (Standing, 1947; Glasø, 1980; Dokla &

				Osman, 1992). AAPRE of 1.87% was reported for this second correlation.
Velarde et al. (1999)	11.6-53.4	N/A	2097	The correlation was derived using the "reduced" PVT properties. AAPRE = 1.74%
Al-Shammasi (2001)	6.0-63.7	Global data	1345	Two different correlations were developed for B_o. The first one, like many B_o correlations, has four dependent variables: R_s,T,γ_g and γ_o , while the other has three correlating variables: R_s,T and γ_o. The author reported AAPRE of 1.806% and 3.033 for his new correlations with four and three variables respectively.
Boukadi et al. (2002)	34.8-136	Middle East	32	22 data set was used to generate the correlation while the remaining was used for testing it. The correlation was developed for $P \leq P_b$. AAPRE = 0.85200.
Dindoruk and Christman (2004)	14.7-40	Gulf of Mexico	99	AAPRE = 2.00%.
Malallah et al. (2006)	14.3-59	World wide	5200 (Another 234 data set for testing)	The authors used a non-parametric optimisation method called alternating conditional expectation (ACE) to develop the correlation. In testing, AAPRE of 1.97% and correlation coefficient of 0.9854 were reported.
Hemmati and Kharrat (2007)	18.8-48.34	Iran	287	AAPRE = 1.08% and correlation coefficient = 0.9929
Sutton (2008)	10.6-63	Worldwide	11960	AAPRE = 2.85. The model was compared with 30 other previous correlations. Four of the previous correlations (Al-Shammasi, 2001; Velarde, Blasingame, & McCain Jr, 1997; F. Farshad, LeBlanc, Garber, & Osorio, 1996) gave better results on the tested data than this model. However, the

				model gave similar or better performance than all other compared correlations for the B_{ob} estimation.
Ikiensikimama and Ajienka (2012)	14.87-53.23	Niger Delta	250	Efficiency of the correlation in calculating hydrocarbon reserves was demonstrated. AAPRE = 1.8552 and correlation coefficient = 0.99
Arabloo et al. (2014)	6.00-56.80	World wide	750+	AAPRE = 2.24% and correlation coefficient= 0.94. The authors performed similar analysis to that done for P_b correlation.

2.1.3. Empirical Correlations for Solution Gas/Oil Ratio

A concise review of some empirical correlations for the prediction of solution gas/oil ratio is presented in Table 2.3. Usually, the correlation of R_s is derived from that of P_b. Hence, R_s is correlated with the following functional forms.

$$R_s = f(P_b, \gamma_g, \gamma_{API}, T) \quad (2.8)$$

$$R_s = f(P_b, \gamma_g, \gamma_o, T) \quad (2.9)$$

Table 2.3. Review of Empirical Correlations for Solution Gas/Oil Ratio

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Vazquez and Beggs (1980)	N/A	Worldwide	5008	A correlation with two different set of coefficients was developed by dividing the data set into two based on $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$.
Obomanu and Okpobiri (1987)	14.98-43(γ_o:0.811-0.966)	Nigeria	503	Data points = 503. In testing, AAPRE = 2.19 and correlation coefficient = 0.97.
Kartoatmodjo and Schmidt (1991)	14.4-59.0	Indonesia, North America, Middle east	5392	Two different R_s correlations were developed for $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$. Similar sensitivity analysis to that

		and Latin America		performed for the P_b prediction was carried out.
Elsharkawy and Alikhan (1997)	19.9-42.76	Kuwait	175	Two different correlations were developed by the authors for $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$. Overall statistical evaluations were given with AAPRE =7.87% and correlation coefficient = 0.9636
Velarde et al. (1999)	11.6-53.4	N/A	2097	The new correlation estimates R_s at and below bubblepoint. AAPRE of 4.73% was reported for the correlation using the value of experimental P_b.
Petrosky Jr and Farshad (1998)	16.3-45.0	Gulf of Mexico	90	The correlation was developed for saturated R_s. AAPRE = 3.8%.
Boukadi et al. (2002)	34.8-112.5	Middle East	32	22 data points were used to generate the correlation while the remaining were used for testing it. The correlation was developed for $P \leq P_b$. AAPRE = 28.6190%.
Valko and McCain (2003)	6.0-36.2	World wide	881	R_s at P_b was estimated in terms of stock tank gas-oil ratio and separator gas-oil ratio. AAPRE = 5.2%.
Dindoruk and Christman (2004)	14.7-40	Gulf of Mexico	104	AAPRE = 7.66%.
Hemmati and Kharrat (2007)	18.8-48.34	Iran	287	AAPRE = 4.07% and correlation coefficient = 0.9911.
Khamehchi et al. (2009)	33.4-124	Unknown	94	12 different data sets were used for testing. Correlation coefficient = 0.95.
Ikiensikimama and Ajienka (2012)	14.87-53.23	Niger Delta	250	Two different correlations were developed for $\gamma_{API} \leq 45$ and $\gamma_{API} > 45$. Efficiency of the correlation in calculating hydrocarbon reserves was

				demonstrated. AAPRE = 13.2386% and correlation coefficient = 0.9236.

2.1.4. Empirical Correlations for Dead Oil Viscosity

A concise review of some empirical correlations for the prediction of dead oil viscosity is presented in Table 2.4. This is one of the most difficult PVT properties. Most evaluated literature correlations have given very high errors when applied on new data sets (Al-Marhoun, 2004) (Oloso et.al. ,2018).

Possible functional forms for developing μ_{od} correlations are the following.

$$\mu_{od} = f(\gamma_{API}, T) \quad (2.10)$$

$$\mu_{od} = f(\gamma_{API}, T, R_s) \quad (2.11)$$

$$\mu_{od} = f(\gamma_{API}, T, P_b) \quad (2.12)$$

$$\mu_{od} = f(\gamma_{API}, T, R_s, P_b) \quad (2.13)$$

Table 2.4. Review of Empirical Correlations for Dead Oil Viscosity

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Beal (1946)	10.1-52.2	US	98	APE = 24.2%. Performance of the correlation was tested by dividing the data sets into different temperature and API ranges.
Beggs and Robinson (1975)	16-58	-	460	APE = -0.64%.The correlating variables were T and γ_o.
Glasø (1980)	20.1-45.8	North Sea	38	The developed μ_{od} correlation is used in correcting for paraffinicity in order to adapt the correlation for different crude oils.
Ng and Egbogah (1983)	5-58	-	394	The authors presented a modified correlation of Beggs and Robinson (1975) and also proposed a new μ_{od} correlation which included pour point temperature as a new correlating

				variable. Based on the 394 data points, the APE of 61%, -5.13% and -4.3% were reported for the original Beggs-Robinson correlation, modified Beggs-Robinson correlation and the newly developed correlation respectively.
Twu (1985)	-4 - 93.1	-	563	AAPE = 7.85%
Bennison (1998)	-	North Sea	16	APE = 13%. The correlation was only recommended for API>20 and *T*>250 ^{o}F.
Kartoatmodjo and Schmidt (1991)	14.4-59.0	Indonesia, North America, Middle east and Latin America	661	The correlation was derived using the functional form of Glaso's correlation. Sensitivity analysis of the correlation to reservoir temperature was performed.
Labedi (1992)	32.2-48.0	Libya	91	APE = -2.61%. When the correlation by Beal (1946) was applied to the data set, a very poor result was observed. This was ascribed to the fact that the Beal's correlation was developed for light California crude oil.
Petrosky Jr and Farshad (1998)	25.4-46.1	Gulf of Mexico	118	AAPRE = 12.38%
De Ghetto et al. (1995)	6-22.3	-	1200	Correlations were developed for heavy and extra heavy crude oils based on the correlation of (Egbogah & Ng, 1990) with APPRE of 30.3% and 41.8% for API<10 and $10 < API \leq 22.3$ respectively.
Elsharkawy and Alikhan (1999)	19.9-48	Middle East	254	AAPRE of 19.3% and correlation coefficient = 0.881
Elsharkawy et al. (2003)		Worldwide	361	An empirical correlation which predicts the entire viscosity curve was derived. It predicts from μ_{od} to the

				undersaturated μ_o. Different forms of the developed correlation were explored and the best result was obtained from the one with 8 variables, having average absolute deviation of 24.3%.
Dindoruk and Christman (2004)	17.4-40	Gulf of Mexico	95	P_b and R_{sb} were introduced in the μ_{od} correlation along with the γ_{API} and T. It was stressed that these additional two properties allow the correlation to capture some of the characteristics of the oil type.
Hossain et al. (2005)	7.1-21.8	-	184	142 and 42 datasets from Chevron and (De Ghetto & Villa, 1994) respectively have been used to develop the correlation. Additional data from (T. Kartoatmodjo & Schmidt, 1994) was also used to evaluate the correlation. Testing the correlation separately on different data sets give AAPRE values between 10.6% and 61% while the overall AAPRE on all the dataset was 30.3%.
Naseri et al. (2005)	17-44	Iran	472	250 data points were used to develop the correlations while 222 data points were used for testing and validation. AAPRE = 7.77% for the original regression data and 15.3% for the testing data set.
Bergman and Sutton (2009)	0.45-135.9	Worldwide	9837	Accuracy of the correlation was reported on different temperature ranges. 1. $\gamma_{API} = 5 - 80$,*T*=35-500^{0}F, AAPRE=16.6%

				2. $\gamma_{API} = 5 - 80$,*T*=35-100^{0}F, AAPRE=18.1% 3. $\gamma_{API} = 5 - 80$,*T*=100-200^{0}F, AAPRE=17.6% 4. $\gamma_{API} = 5 - 80$,*T*=200-300^{0}F, AAPRE=15.2%
El-hoshoudy et al. (2013)	21-52	Egypt	1000	AAPRE = 9.8855% and correlation coefficient = 0.9285.
Alomair et al. (2014)	10-20	Kuwait	374/118	374 data points were used for developing the correlation while 118 data point were used for testing. For the training data set AAPRE = 25.29% while for the testing data set AAPRE = 28.08%.

2.1.5. Empirical Correlations for Saturated Oil Viscosity

A review of some empirical correlations for the prediction of saturated oil viscosity is presented in Table 2.5. Common functional forms found in the literature are the following.

$$\mu_{ob} = f(\gamma_g, R_s, \gamma_o, T) \quad (2.14)$$

$$\mu_{ob} = f(\gamma_g, R_s, \gamma_{API}, T) \quad (2.15)$$

$$\mu_{ob} = f(\gamma_{API}, \mu_{od}, P_b) \quad (2.16)$$

$$\mu_{ob} = f(\mu_{od}, R_s) \quad (2.17)$$

$$\mu_{ob} = f(\mu_{od}, P_b) \quad (2.18)$$

Table 2.5. Review of Empirical Correlations for Saturated Oil Viscosity

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Beal (1946)	15.8-45.7	US	351	A graphical presentation of gas-free crude oil was presented. It was

				stated that the amount of dissolved gas has more impact on μ_{ob} than pressure. The correlation was evaluated with APE = 13.4%.
Chew and Connally (1959)	NA	US	457	The performance of the correlation was examined by using the confidence limit.
Beggs and Robinson (1975)	16-58	-	2073	APE = -1.83%. The live oil viscosity was correlated as a function of μ_{od} and R_s.
Khan et al. (1987)	14.3-44.6	Saudi Arabia	150/1691	A total of 150 and 1691 data points were used to develop viscosity correlations at and below P_b with AAPRE of 12.148% and 5.157% respectively. The reported correlation coefficients were 0.953 and 0.994 for viscosity correlations at and below P_b respectively.
Kartoatmodjo and Schmidt (1991)	14.4-59.0	Indonesia, North America, Middle east and Latin America	5321	A total of 5321 data points were used to develop a μ_{ob} correlation. Similar sensitivity analysis as performed for the μ_{od} prediction was carried out.
Khan et al. (1987)	21-49	Canada and Middle East	459	AAPRE = 4.91% and correlation coefficient = 0.9979.
Labedi (1992)	32.2-48	Libya	91	μ_{ob} correlation was developed with the following independent variables: γ_{API}, μ_{od} and P_b and the correlation's APE was -2.38%. Another correlation was developed for μ_o below P_b. For the μ_o correlation below P_b, a linear relationship between μ_o and P was established for the pressure range

				$P_b > \mathrm{P} > 0.15P_b$. The slope of the linear relationship between μ_o and P was correlated with APE = 3.5%.
Petrosky Jr and Farshad (1995)	25.4-46.1	Gulf of Mexico	864	AAPRE = 14.47%.
De Ghetto et al. (1995)	6-22.3	-	1200	Correlations were developed for heavy and extra heavy crude oils based on correlation of[53] with APPRE of 14.7% and 16.1% for API<10 and $10 < API \leq 22.3$ respectively.
Almehaideb (1997)	30.9-48.6	UAE	57	For μ_{ob} correlation, the reported AAPRE = 13% which is smaller than the results from all other compared correlations. Correlation coefficient = 0.9691.
Hanafy et al. (1997)	17.8-47.7	Egypt	324	AAPRE = 19.1% and correlation coefficients = 0.91.
Elsharkawy and Alikhan (1999)	19.9-48	Middle East	254	AAPRE of 18.6% and correlation coefficients = 0.978.
Boukadi et al. (2002)	34.8-136	Middle East	32	22 data points were used to generate the correlation while the remaining was used for testing it. The correlation was developed for $P \leq P_b$. AAPRE = 35.5560%.
Dindoruk and Christman (2004)	17.4-40	Gulf of Mexico	95	AAPRE = 13.2%.
Naseri et al. (2005)	17-44	Iran	472	AAPRE = 16.4% for the original regression data and 26.3% for the testing data set.
Hossain et al. (2005)	7.1-22.3	-	415	The saturated viscosity was developed as a function of μ_{od} and R_s. The new correlation gave AAPRE of 53.2, 46.4 and 26.5%

				when applied to datasets from Kartoatmodjo and Schmidt (1994), De Ghetto et al. (1995) and Chevron respectively.
Bergman and Sutton (2007)	6.0-61.7	Worldwide	12474	The author proposed a new correlation as a result of the observed inconsistency in the behaviour of all the evaluated correlations. For the new correlation, AAPRE = 12.4%.
Khamehchi et al. (2009)	33.4-124	-	94	Correlation coefficient = 0.98.
El-hoshoudy et al. (2013)	21-52	Egypt	1000	AAPRE = 11.2281% and correlation coefficient = 0.9493.
Ghorbani et al. (2016)	21.55–30.62	Iran	600	The developed correlations are fairly large with 36 and 42 coefficients for viscosity below bubblepoint and saturated viscosity respectively. It could be a bit cumbersome for field application. APPRE of 3.77% and 0.01058% were reported for viscosity below bubblepoint and saturated viscosity respectively.

2.1.6. Empirical Correlations for Undersaturated Oil Viscosity

A concise review of some empirical correlations for the prediction of undersaturated oil viscosity is presented in Table 2.6. It is observed that most authors have reported good results for their correlations of μ_{oa} and different evaluations of some these correlations have also shown acceptable results on new data sets (Al-Marhoun, 2004; Mahmood & Al-Marhoun, 1996). Different sets of input correlating variables have been used. Some of the common functional forms include the following.

$$\mu_{oa} = f(\mu_{ob}, P_b, P) \quad (2.19)$$

$$\mu_{oa} = f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API}) \quad (2.20)$$

$$\mu_{oa} = f(\mu_{ob}, \mu_{od} P_b, P) \quad (2.21)$$

Table 2.6. Review of Empirical Correlations for Prediction of Undersaturated Oil Viscosity

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Beal (1946)	-	US	52	The correlation was evaluated with APE = 2.7%.
Vazquez and Beggs (1980)	-	Worldwide	6000	The correlation was developed without γ_{API} grouping unlike other correlations developed in the same paper.
Khan et al. (1987)	14.3-44.6	Saudi Arabia	1503	AAPRE = 1.915% and correlation coefficients = 0.999.
Kartoatmodjo and Schmidt (1991)	14.4-59.0	Indonesia, North America, Middle east and Latin America	3588	Sensitivity analysis based on grouping of P_b and R_s was carried out.
Labedi (1992)	32.2-48	Libya	91	The general equation of a straight line was adopted in developing the correlation. The slope of the equation is the parameter that was correlated with APE = -3.1%.
Petrosky Jr and Farshad (1995)	25.4-46.1	Gulf of Mexico	404	AAPRE = 2.91%.
De Ghetto et al. (1995)	6-22.3	-	1200	Correlations were developed for heavy and extra heavy crude oils based on the correlations of Labedi (1992) for API<10 and Kartoatmodjo & Schmidt (1994) for ($10 < API \leq$

				22.3) with APPRE of 12.3% and 10.1% respectively.
Almehaideb (1997)	30.9-48.6	UAE	328	For μ_o correlation, AAPRE = 2.885% while the compared correlation of Vazquez and Beggs (1980) gave APPRE of 8.58%.
Elsharkawy and Alikhan (1999)	19.9-48	Middle East	254	AAPRE = 4.9% and correlation coefficient = 0.972.
Dindoruk and Christman (2004)	17.4-40	Gulf of Mexico	93	AAPRE = 5.99%
Naseri et al. (2005)	17-44	Iran	472	AAPRE = 2.12% for the original regression data and 3.62% for the testing data set.
Hossain et al. (2005)	7.1-22.3	-	390	AAPRE of 31.2%, 38.9% and 56.3% were got for the data sets of Chevron and other previous works (De Ghetto & Villa 1994; Kartoatmodjo & Schmidt 1994).
Ghorbani et al., (2016)	21.55–37.62	Iran	600	The developed correlation has 36 coefficients which is very rare in the literature. AAPRE of 0.268% was reported for the correlation.

2.1.7. Empirical Correlations for Isothermal Oil Compressibility

A concise review of some empirical correlations for the prediction of isothermal oil compressibility is presented in Table 2.7. Most of the C_o correlations are for saturated ($P = P_b$) and undersaturated reservoir conditions ($P > P_b$). Very few correlations have been developed the reservoir condition below P_b (Muhammad Ali Al-Marhoun, 2003). Common functional forms that used for developing these correlations below, followed by a concise review of some of the correlations in the literature.

$$C_o = f(R_s, T, \gamma_g, \gamma_o, P) \quad (2.22)$$

$$C_o = f(R_s, T, \gamma_g, P) \quad (2.23)$$

$$C_o = f(R_s, T, \gamma_g, \gamma_{API}, P) \quad (2.24)$$

$$C_o = f(\gamma_o) \quad (2.25)$$

$$C_o = f(P_b) \quad (2.26)$$

Table 2.7. Review of Empirical Correlations for Isothermal Oil Compressibility

Authors	γ_{API} Range	Region of Data Source	Data Points	Comment
Calhoun Jr (1947)	-	-	-	Graphical correlation was developed for C_o.
Vazquez and Beggs (1980)	N/A	Worldwide	4036	A correlation with two different set of coefficients was developed by dividing the data set into two based on $\gamma_{API} \leq 30$ and $\gamma_{API} > 30$. The correlation is used in calculating the undersaturated B_o
Kartoatmodjo and Schmidt (1991)	14.4-59.0	Indonesia, North America, Middle east and Latin America	3588	Similar sensitivity analysis to that performed for the P_b prediction was carried out.
Al-Marhoun (1992)	10.4-49.2	Worldwide	2000	The developed correlation for undersaturated C_o is utilised in the correlation for undersaturated B_o
De Ghetto et al. (1995)	6-22.3	-	1200	Correlation were developed for heavy and extra heavy crude oils based on the correlation of[13] with APPRE of 12.3% and 10.1% for API<10 and $10 < API \leq 22.3$ respectively.
Almehaideb (1997)	30.9-48.6	UAE	244	Reported APPRE = 9.88% while the compared correlation of Vazquez and Beggs (1980) gave APPRE of 25.71%. The author noted that Vazquez and Beggs[13] correlation significantly underestimated the C_o for highly compressible UAE crude oils.

Elsharkawy & Alikhan (1997)	19.9-42.76	Kuwait	423	The undersaturated C_o was correlated as a function of R_s,T and P. Both γ_o and γ_g were excluded unlike similar correlations by Vazquez and Beggs (1980) and Kartoatmodjo and Schmidt (1991) which included the two parameters. AAPRE = 15.23% and correlation coefficient = 0.8108.
Hanafy et al. (1997)	17.8-47.7	Egypt	324	AAPRE = 10.84% and correlation coefficients = 0.92.
Petrosky Jr and Farshad (1998)	16.3-45.0	Gulf of Mexico	304	AAPRE = 6.66%.
Dindoruk and Christman (2004)	14.7-40	Gulf of Mexico	-	No metric or statistical evaluation of the developed correlations were given in the paper.
Sutton (2008)	10.6-63	Worldwide	11960	Correlation using hyperbolic function was derived to estimate the instantaneous C_o. Also, an equation was derived to estimate the average C_o. AAPRE of 11.7% and 12.05% were reported for testing the instantaneous and the average models respectively.
Al-Marhoun (2003)	-	Middle East	3412	The correlation was developed using 3412 data points while additionally four different data sets of 495, 246, 182 and 182 data points from Canada, Pakistan, Yemen and Vasquez's thesis respectively have been used for validation. For the developed correlation, AAPRE of 5.56%, 10.09%, 16.30, 17.24% and 23.43% were reported for data sets respectively.
Al-Marhoun (2009)	-	-	-	The paper criticised the current definition of C_o below bubblepoint. New equations for the saturated and undersaturated C_o were derived.

2.2. API Gravity Classification of Crude Oils

Crude oil is classified based on API gravity (γ_{API}) to determine its heaviness which consequently determines its marketability. Table 2.8 shows a typical oil classification based on API gravity (Dandekar, 2013; De Ghetto, Paone, & Villa, 1995). Knowledge of the API gravity and other PVT properties such as bubblepoint pressure (P_b), oil formation volume factor (B_o) and oil viscosity are important for determining future production or oil reserves from petroleum wells.

Table 2.8**.** Crude oil Classification based on API

Classification	API Range
Light	API > 31.1
Medium	$22.3 \leq$ API ≤ 31.1
Heavy	API < 22.3
Extra Heavy	API < 10.0

γ_{API} is calculated using the specific gravity of an oil (γ_o) which is the ratio of oil density to that of water. Specific gravity for API is normally determined at 60 degrees Fahrenheit. It is thus given as:

$$\gamma_{API} = \frac{141.5}{\gamma_o} - 131.5 \qquad (2.27)$$

Chapter 3. Machine Learning Techniques: Brief Introduction and Application to Reservoir Fluid Characterisation

3.1. Overview of Machine Learning

There is no unanimous definition of Machine Learning among the experts. ML is basically the prediction of the future with (supervised learning) or without, (unsupervised learning), prior knowledge of the past. Learning involves creation of a system which applies past experience to analogous new situations. ML can also be described as the process of writing computer programs to optimise a performance criterion using example data or past experience (Alpaydin, 2014). Learning can be in or through many forms; it can be through new knowledge acquisition, cognitive skills acquisition, effective representation of new knowledge or new fact discovery through observation and experimentation (Carbonell, Michalski, & Mitchell, 1983).

In ML, a model is built which utilises some parameters that are optimised using the training data or past experience. The expected outputs from the learning model may be known (predictive/supervised problems) or unknown (descriptive/unsupervised problems). Also, the target output from the ML model may be continuous (classification problems) or discreet (regression problem).

Some of the notable ML techniques that are connected to this work are briefly explored.

3.2. Artificial Neural Networks

Neural Networks or Artificial Neural Networks (ANNs), to be more precise, are endowed with some unique attributes (just like other ML techniques): universal approximation (input-output mapping), the ability to learn from data and adapt to their environment and the ability to invoke weak assumptions about the underlying physical phenomena responsible for the generation of the input data. A neural Network is a massively parallel distributed processor that has a natural propensity for storing experimental knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.

2. Interneuron connection strengths known as synaptic weights are used to store knowledge.

The procedure to perform the learning process is called a learning algorithm. The synaptic weights of the network are modified in an orderly fashion so as to attain a desired design objective (Príncipe, Euliano, & Lefebvre, 2000). A neural network is made up of an interconnection of neurons.

3.2.1. Model of a Neuron

A neuron is an information-processing unit that is fundamental to the operation of an ANN. In essence, neurons are like signal processing elements for the neural network. The model of a neuron is shown in Figure 3.1.The three basic elements of a neuron are described thus.

1. A set of *synapses or connecting links*, each of which is characterised by a weight or strength of its own. Specifically, a signal x_j at the input of synapse *j* connected to neuron *k* is multiplied by the synaptic weight w_{kj} .
2. An *adder* for summing the input signals, weighted by the respective synapses of the neuron.
3. An *activation function* for limiting the amplitude of the output of a neuron. The activation function is also referred to as *squashing function*. Typically, the normalised amplitude range of the output of a neuron is written as the closed unit interval [0 1] or alternatively [-1 1].

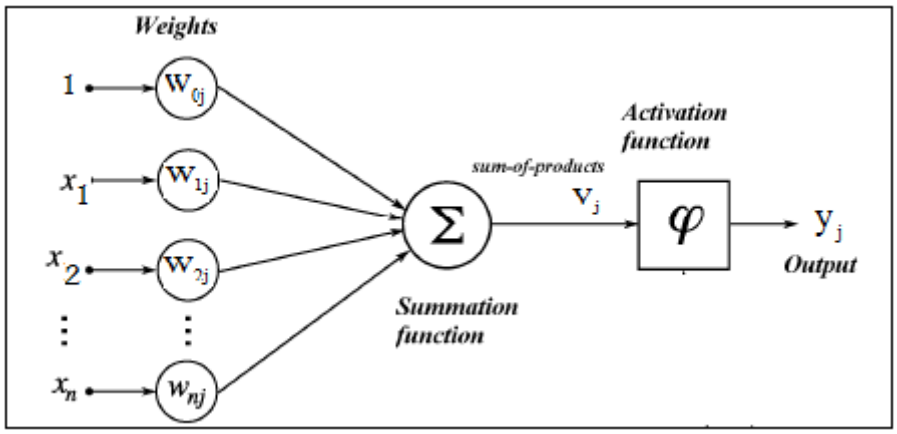


Figure 3.1. Nonlinear Model of a Neuron

Mathematically,

$$V_j = \sum_{k=1}^{n} W_{kj} X_k + W_{oj} \tag{3.1}$$

$$Y_j = \varphi(V_j) \tag{3.2}$$

Where $X_1, X_2, \ldots\ldots\ldots, X_n$ are the input signals. $W_{1j}, W_{2j,\ldots\ldots\ldots\ldots}, W_{nj}$ are the synaptic weights of neuron j, W_{oj} is the bias, V_j is the linear combiner output, $f(\cdot)$ is an activation function and Y_j is the output signal of the neuron.

3.3. Support Vector Machines

SVM originates from statistical learning which is used for learning separating functions in pattern recognition (classification) tasks or for performing functional estimation in regression problems. SVM was first developed for the restricted case of separating training data without errors and this was extended to cover when separation without error on the training vectors is impossible, hence, culminating in a generalised learning algorithm (Cortes & Vapnik, 1995).

In simple pattern recognition problems, SVM uses a linear separating hyperplane to create a classifier with a maximal margin. Invariably, the learning problem is cast as a constrained nonlinear optimisation problem. In this, the cost function is quadratic and the constraints are linear (Kecman, 2001).

On the other hand, for a regression problem or where the input classes cannot be linearly separated in the original input space, SVM first nonlinearly transforms the original input space into a higher dimensional space where a maximal separating hyperplane is constructed (Cortes & Vapnik, 1995). The main aim is to minimise the error in the training with maximisation of the margin in the hyperplane with appreciable error tolerance to enhance model generalisation (Kecman, 2001).

Figure 3.2 shows how a margin is created between two sets of data in a classification problem.

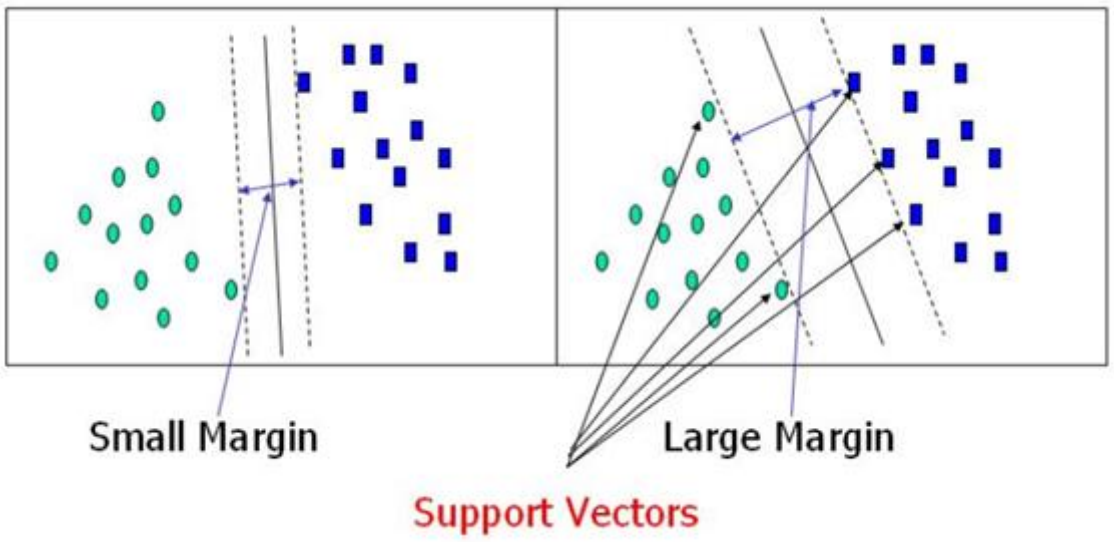


Figure 3.2. Creation of margins between two datasets by support vectors (Adapted from (Cortes & Vapnik, 1995))

Mathematically, since the main idea is to optimise the margin then the quadratic optimisation problem becomes

$$\min_{w} \tfrac{1}{2} W^T W \tag{3.3}$$

$$s.t. \begin{Bmatrix} y_i - (W^T \phi(X) + b) \le \varepsilon \\ (W^T \phi(X) + b) - y_i \le \varepsilon \end{Bmatrix} \tag{3.4}$$

Where $\phi(X)$ is the kernel function, W is the margin and the pair (x_i, y_i) is the training set. Then we add a bound in order to set the tolerance on errors number that can be committed:

$$\min_{W,b} \frac{1}{2} W^T W + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{3.5}$$

$$s.t. \begin{Bmatrix} y_i - (W^T \phi(X) + b) \le \varepsilon + \xi_i \\ (W^T \phi(X) + b) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1, 2, ..., l \end{Bmatrix} \tag{3.6}$$

Once it is trained, SVM will generate predictions using the following formula:

$$f(X) = \sum_{i=1}^{l} \theta_i \phi(X, X_i) + b \tag{3.7}$$

For the kernel, possible options are functions such as: Gaussian, polynomial, radial basis and wavelet. The kernel plays the most important role in determining the accuracy of SVM prediction. In his renown book, Vapnik (1995) introduced support vector regression (SVR) machine which is the first standard version of SVM that addresses regression problems.

3.4. Decision Trees

A decision tree is a ML technique for building prediction models from data. It is a hierarchical data structure which implements divide-and-conquer strategy (Alpaydin, 2014). It is generally a non-parametric method that is used for both classification and regression problems. Decision trees are non-parametric because they do not assume any distributional properties about the data (Lior Rokach, 2015). A decision tree creates a sequence of recursive splits in a number of steps. A decision tree composes of decision nodes and terminal leaves where each node *n* implements a function $f_n(x)$, giving discrete outcomes with branches.

A decision tree is of two types: classification and regression trees. Classification trees apply to problems where the output data is discreet while RTs apply to problems where the output is

continuous. Recursive partitioning of the data space is usually used to fit a simple prediction model in each partition to build a decision tree model.

3.4.1. Classification Tree

A classification tree is a type of decision tree where the outcomes of each tree node is a label i.e. non-numeric. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost (Loh, 2011).

In a classification problem, we have for instance *p* predictor variables $\{X_1, X_2, \dots, X_p\}$ with a training sample of *n* observations for class variable *Y* that takes values 1, 2, ... , *k.* The goal is to find a model for predicting the values of *Y* for new values of *X*. Theoretically, the solution is simply a partition of the *X* space into *k* disjoint sets, $A_1, A_2, \dots, A_k$, such that the predicted value of Y is j if X belongs to A_j, for *j*= 1, 2, …, *k* (Loh, 2011).

The first classification algorithm in literature is THAID (Messenger & Mandell, 1972). THAID uses a measure of node impurity based on the distribution of the observed *Y* values in the node, splitting a node by exhaustively searching over all *X* and *S* for the split $\{X \in S\}$ which minimises the total impurity of its two child nodes. The algorithm for tree construction by exhaustive search is shown in Figure 3.3.

1. Start at the root node
2. For each X, find the set *S* that minimises the sum of the node impurities in the two child nodes and choose the split $\{X^* \in S^*\}$
3. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

Figure 3.3. Pseudo-code for tree construction by exhaustive search

Other popular algorithms in the literature for training classification trees include Classification And Regression Trees (CART) (Breiman, Friedman, Stone, & Olshen, 1984), C4.5 (Quinlan, 2014), Fast and Accurate Classification Tree (FACT) (Loh & Vanichsetakul, 1988), Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE) (Kim & Loh, 2001) (Kim & Loh, 2003), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) (Loh, 2009) and Quick, Unbiased and Efficient Statistical Tree (QUEST) (Loh & Shih, 1997), among others.

A very simple example of a classification tree is illustrated in Figure 3.5. This is deduced from the graph in Figure 3.4.

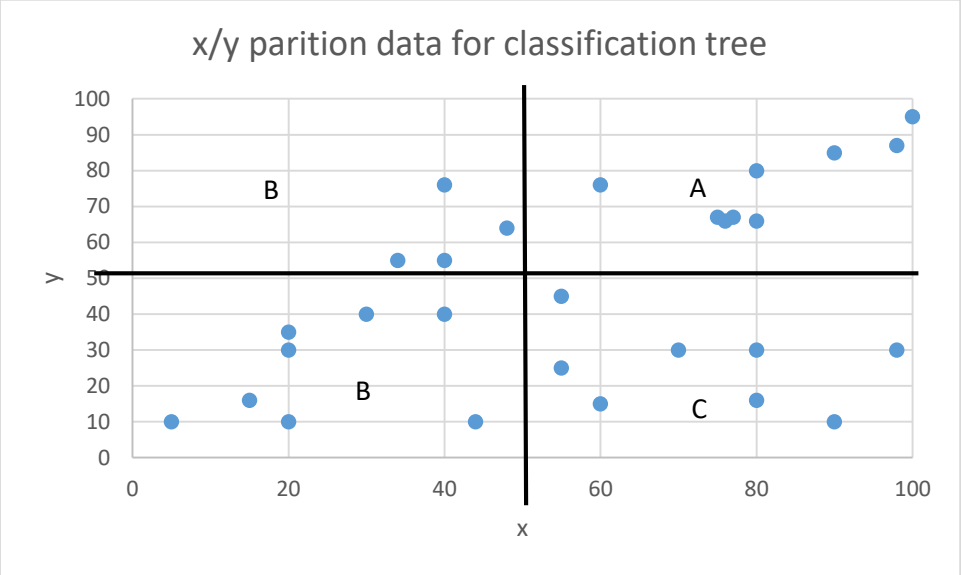


Figure 3.4. Simple partitions of data for classification tree illustration

In Figure 3.4, a set of data has been partitioned for illustration of a simple classification tree splitting process. The resulting decision tree where the data has been classed into three different groups based on simple "IF-THEN" statements is shown in Figure 3.5.

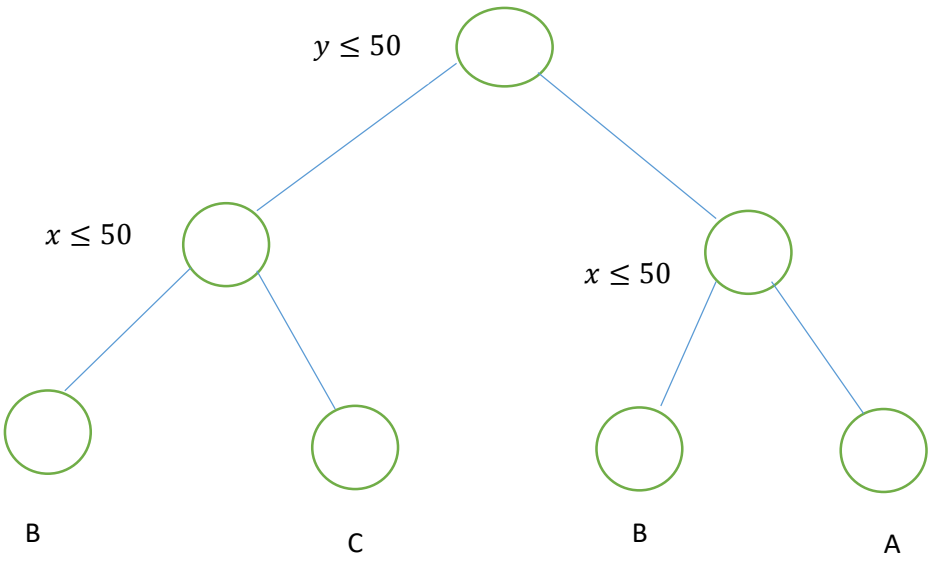


Figure 3.5. A Classification tree model for the partitions in Figure 3.4.

3.4.2. Regression Tree

A RT is a variant of decision tree which maps the input space into a real valued domain. In RT, the target output consists of numeric or continuous values (real numbers) (Lior Rokach, 2015). The first RT algorithm is Automatic Interaction Detection (AID) (Morgan & Sonquist, 1963), predating THAID algorithm. Another influential RT algorithm is the CART. Both AID and CART algorithms follow the Algorithm in Figure 3.3, where the node impurity is the sum of squared deviations about the mean and the node predicting the sample mean of the target.

Both AID and CART construct piecewise constant RTs by using the node mean of Y as predicted value and the sum of squared deviations as node impurity function. Subsequently developed RT algorithms fall under one of two directions (W. Loh, 2014):

(i) Piecewise linear or higher order least-squares models

(ii) Piecewise constant or linear models with other loss functions.

Other RT algorithms include M5 (Quinlan, 1992) and its variant M5' (Wang & Witten, 1996) which is less computationally intensive; Smoothed and Unsmoothed Piecewise Polynomial Regression Trees (SUPPORT) (Chaudhuri, Huang, Loh, & Yao, 1994); GUIDE (W.-Y. Loh, 2002); CTREE (Hothorn, Hornik, & Zeileis, 2006) and Regression Trunk Approach (RTA) (Dusseldorp & Meulman, 2004).

3.5. Functional Networks

Functional networks (FNs) were introduced as a powerful alternative to neural networks (Castillo, 1998; Castillo et al., 2000). Unlike neural networks, functional networks have the advantage that they use domain knowledge in addition to data knowledge. The network initial topology can be derived based on the modelling of the properties of the real world. Once this topology is available, functional equations allow one to obtain a much simpler equivalent topology.

Simplified general topologies for ANN and FN are shown in Figures 3.6 and 3.7 respectively. In these figures, X_1, X_2 and X_3 are the inputs into the network. X_4 and X_5 are the outputs of the hidden layer. W_{mn} (*m*= 4, 5, 6; *n*= 1, 2, 3, 4, 5) are the weights while Y is the output in both cases.

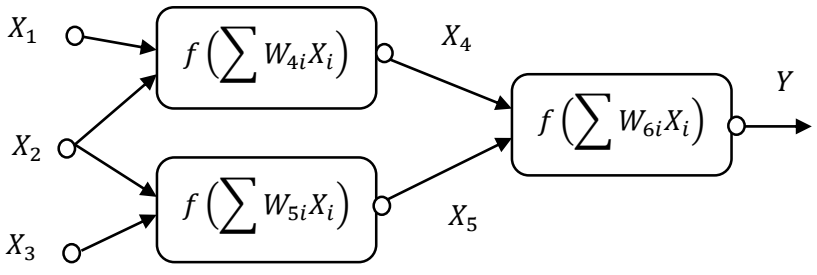


Figure 3.6. A Standard Neural Network

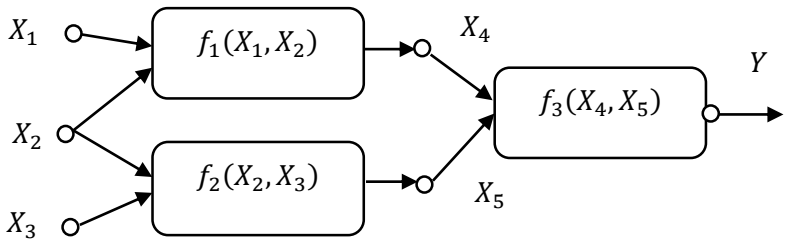


Figure 3.7. A Standard Functional Network

There are some quite significant differences between ANN and FN. Notably, the functions in FN are truly learned during the structural learning unlike the ANN where neuron functions are assumed to be fixed and known and only the weights are learned. The implemented FN is shortly described below.

Given a data set $\{x_{ij}|y_i; i = 1,2, \ldots . n \ \& \ j = 1,2,3,4\}$ where x_{ij} are the predictors and y_i is the output. Mathematically, the relationship can be given by

$$Y = f(X_1, X_2, X_3, X_4) \tag{3.8}$$

Note that *j* refers to the number of inputs which is 4 here. The general form of a functional network that learns from the data can be given as follows (Castillo et. al., 2000).

$$y_i = \sum_{j=1}^{p} \sum_{r=1}^{m} C_r \varphi_r(x_{ij}), \quad i = 1,2, \ldots, n \tag{3.9}$$

where φ_r are the linear independent functions which are used to learn the coefficients C_r. Some possible function for φ_r are:

(1). Polynomial function:

$$\varphi = \{1, x, x^2, \ldots, x^m\} \tag{3.10}$$

(2). Exponential Function:

$$\varphi = \{1, e^x, e^{-x}, \ldots, e^{mx}, e^{-mx}\} \tag{3.11}$$

(3). Fourier Function:

$$\varphi = \{1, \sin(x), \cos(x) \ldots, \sin(mx), \cos(mx)\} \tag{3.12}$$

4). Logarithm Function:

$$\varphi = \{1, \log(x+2), \log(x+3), \ldots, \log(x+m)\} \quad (3.13)$$

The aim is to get $\hat{Y}$ which is an estimate of Y such that the square of the error is minimised. That is;

$$\min\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right\}. \quad (3.14)$$

Hence, the aim is to produce an estimate $\hat{Y}$ that gives minimal error ε which can be represented as:

$$\varepsilon = \min(Y - \widehat{Y}). \quad (3.15)$$

This final equation can be solved using least square optimisation. Several useful analyses and applications of FNs can be found in the literature (Elsebakhi et al., 2015; Asafa et al., 2015).

3.6. Ensemble Machine Learning System

Ensemble ML is a combination of multiple base models-classifiers or regressors. Each base model covers a different part of the input space or the complete input space. Though there is no definitive taxonomy for building the ensemble models, some successful approaches and methodologies have been widely adopted (Dietterich, 2000). Popular among these methods are bagging and boosting. Another notable method, mainly for classification problem is ADABOOST (Freund & Schapire, 1996).

In ensemble ML, each base model is usually trained on a slightly different training set and the predictions from all the models are combined with the goal of producing a better and more accurate output than the individual base models. The ensemble models aim to reduce the expected errors in the predicted target output values. Expected error consists of the prediction bias and the variance (Friedman, Hastie, & Tibshirani, 2001). In other words, it is the expected difference between the estimated function and the true function.

The prediction error from high bias and low variance in some base models (causing under fitting of the training data sets) or low bias and high variance in other base models (causing over fitting of the training data sets) can be reduced when different base models are combined.

Bagging (Breiman, 1996), which is also known as bootstrap aggregation involves training multiple models with training sets of data randomly drawn with replacement from the base training datasets. The training datasets for the base models are called bootstraps. Hence, bagging involves training different models with different samples and usually predictions are obtained by averaging the results of the different base models for a regression problem.

Boosting involves training and improving a weak learning algorithm into a strong one (Schapire, 1990). In boosting, the training dataset for each subsequent model increasingly focuses on instances wrongly predicted by the previous weaker model. ADABOOST (adaptive boosting algorithm) is one of the most used boosting algorithms which automatically adapts to the data given to it.

Generic representation of ensemble system is shown in Figure 3.8.

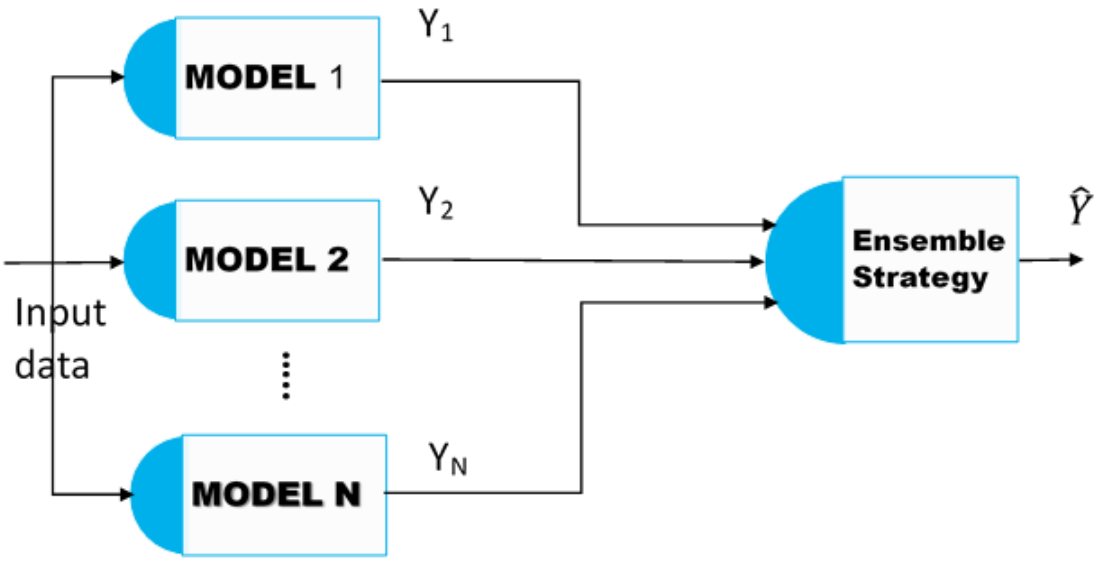


Figure 3.8. General representation of an ensemble system

3.7. Overview of Machine Learning Techniques for Characterisation of Reservoir Fluid Properties

The potential applications of machine learning (ML) or soft computing techniques in general have been noticed by some researchers for several years (Ali, 1994; Mohaghegh, 2000). A relatively small number of studies have been carried out using ML techniques in the petroleum industry, including modelling of PVT properties using neural networks. Although use of ML has not been widely adopted in the petroleum industry, a number of techniques other than the commonly used neural networks have been considered.

Gharbi & Elsharkawy, (1997) developed artificial neural network models to predict P_b and B_{ob} based on 498 experimentally obtained data sets. Performance of their models for these two PVT properties were compared with the performances of the correlations developed by (Al-Marhoun, 1988), (Standing, 1947) and (Glasø, 1980). For P_b, their ANN model gives AAPRE of 6.89% compared to 18.10%, 20.64% and 18.78% for (Al-Marhoun, 1988), (Standing, 1947) and

(Glasø, 1980) respectively, based on evaluation using testing data. The CC, based on evaluation of the testing data, for the P_b ANN model was 0.962 while the CCs given by (Al-Marhoun, 1988), (Standing, 1947) and (Glasø, 1980) are 0.458, 0.574 and 0.787 respectively, For B_{ob} testing , their ANN gives AAPRE of 2.79% while the CCs of (Al-Marhoun, 1988), (Standing, 1947) and (Glasø, 1980) give 4.22%, 4.42% and 4.97% respectively. The correlations for the B_{ob} testing are 0.979, 0.925, 0.923 and 0.911 for their ANN models, (Al-Marhoun, 1988), (Standing, 1947) and (Glasø, 1980) respectively.

Elsharkawy (1998) developed two RBF networks with the architectures 4-100-100-4 and 4-100-100-2, to predict some PVT properties. The first network with the structure 4-100-100-4 was used to predict R_s,B_o, μ_o and γ_o with with testing AAPREs of 4.53%, 0.53%, 8.72% and 0.4% respectively, while the second network was used to predict undersaturated C_o and γ_g with testing AAPREs of 5.98% and 3.03% respectively.

Gharbi et al. (1999) developed neural network models to predict P_b and B_{ob} using 5200 PVT data points for training the networks and some additional 234 data points for testing. They compared their P_b and B_{ob} ANN models with 8 and 12 empirical correlations respectively. Their P_b ANN model gave AAPRE of 6.48% and CC of 0.9891 for testing. Of all the compared P_b empirical correlations, Standing (1947) gave the best CC (0.9652) with AAPRE of 16.69% .For B_{ob} correlation testing, their ANN model gave AAPRE of 1.97% and CC of 0.9875. They recorded the best empirical B_{ob} correlation for Kartoatmodjo & Schmidt (1994) with CC of 0.9794 and AAPRE of 1.86%.

Varotsis et. al (1999) developed ANN models using over 650 reservoir fluids to predict both oil and gas condensate PVT properties. The predicted oil properties include:B_o, C, R_s, ρ_o, μ_o, z factor at 90% of P_b and liberated gas relative density, while the predicted gas condensate properties are : constant mass study (CMS) relative volume(%), CMS- maximum condensation (%), constant volume depletion study (CVDS)- maximum condensation (%), gas relative density and z factor(%). They used two approaches called “the curve shape approach” and the “reference value approach”. It was reported that most of the PVT properties were estimated with low mean relative error between 0.5 and 2.5%.

Osman et al. (2001) developed a feed forward neural network model to predict B_{ob} using 803 data points. 402 and 201 data points were used to develop and test the model respectively. In testing, their ANN model gave AAPRE of 1.7886% and CC of 0.9878 for B_{ob} prediction. Among all the evaluated empirical correlations, the best correlation result was obtained from the correlation of Vazquez & Beggs (1980) with a CC of 0.9842 and AAPRE of 2.9755%.

Al-Shammasi (2001) performed evaluation of 13 different empirical correlations for P_b and B_{ob}. After cleansing the initial data sets, 1,243 and 1,345 data points were used for used for the study. Based on these data sets, the coefficients of the evaluated correlations were re-calculated. Also, he developed new empirical correlations for both P_b and B_{ob}. Among the compared correlations, his new correlations gave the best results with CC of 0.9987 and AAPRE of 17.85% for the P_b prediction, and AAPRE of 1.806% for the B_{ob} prediction. He also developed ANN models to predict P_b and B_{ob}. For the P_b ANN model, he used 1106 and 137 data points for training and testing respectively, while for the B_{ob} ANN model, he used 1165 and 180 data points for training and testing respectively. He reported AAPRE of 19.86% on testing data sets for the P_b ANN model but he only reported the AAPRE (11.68%) on the training data sets for B_{ob} ANN model. He emphasised that the numerical correlations outperformed the developed ANN models for the two PVT properties.

Al-Marhoun & Osman (2002) developed ANN models to predict P_b and B_{ob} using 283 data points from Saudi oil fields, with 142, 71 and 70 data points for training, cross-validation and testing respectively. The architectures of the ANN used were 4-7-1 and 4-8-1 for P_b and B_{ob} respectively. They compared the results of their models with some existing numerical models. For B_{ob} prediction, their ANN model gives the best result with CC of 0.9989 and AAPRE of 0.5116%. Among all the compared empirical equations, Al-Marhoun (1992) gave the best result with AAPRE of 0.845%. With respect to the P_b prediction, their ANN model also outperformed all the compared empirical equations with CC of 0.9965 and AAPRE of 5.8915%. Among the compared empirical equations in this case, Al-Marhoun (1988) gave the best result with AAPRE of 8.1028%. The authors indicated that the two empirical models which have given better predictions than other compared empirical equations were developed using Saudi oil fields samples like their study. According to them, this is an evidence that regional correlations for predicting PVT properties are better rather than universal ones suggested by others.

Osman & Abdel-Aal (2002) used "abductive network" which is based on group method of data handling to predict P_b and B_{ob}. In building the models, they used 198 and 85 data points for training and testing respectively. In testing the P_b model, they reported CC of 0.9898 and AAPRE of 5.62% for P_b, while CC of 0.9959 and AAPRE of 0.86% were reported for testing the B_{ob} model.

Goda et.al. (2003) designed four layers ANN architectures as 4-10-10-1 and 5-8-8-1 to predict P_b and B_{ob} respectively. They used the predicted P_b as one of the inputs to the ANN model for predicting the B_{ob}, rather than the experimentally obtained values. They used 160 and 20 data points for training and testing the models. The data points were from Middle East oil samples. CC

of 0.998 and AAPRE of 3.0704% were reported in testing for the P_b ANN model. For their B_{ob} ANN model testing, they reported a CC of 0.9936 and AARE of 0.368%.

Osman & Al-Marhoun (2005) developed radial basis functions (RBF) networks to predict PVT properties of oil field brines. The first model has 3-38-3 structure where the input neurons indicate reservoir temperature, pressure and salinity while the three output neuron indicate the predicted values for C, B_o, and ρ. The second model is a backpropagation ANN with the structure 2-2-1. The inputs are temperature and salinity while the only output is the predicted viscosity. They used 780 and 260 data points for training and testing respectively. In testing the RBF models, they reported AAPREs of 0.0981%, 1.0643%, 0.1305% and 1.908% for the brine B, C, ρ and μ. According to the authors, the ANN models gave better performance than the compared empirical correlations.

Ayoub et.al. (2007) evaluated two famous empirical correlations for evaluating undersaturated viscosity. They also developed an ANN model as an alternative because of the poor performance of those correlations in their study. The ANN model was built using feed forward neural network with 7-8-8-1 structure. They used 99 data points from Pakistani crude oil with almost half of the data points for training, one quarter for evaluation and remaining for testing. In testing their model, they reported CC of 0.9931.

El-Sebakhy et al. (2007) used SVM to develop models for predicting P_b and B_{ob} using 782 data points. These data points were divided into three: 382 were used for training and each set of 200 data points were used for validation and testing. They reported to have carried out quality checks to remove redundant data and un-useful observations. They compared their results with those of ANN model and three empirical correlations (Al-Marhoun, 1988; Glasø, 1980 and Standing, 1962). They reported that SVM had the best performance.

Hajizadeh (2007) used genetic algorithm (GA) to predict the viscosity of Iranian black oil. In testing the model, a CC of 0.99742 was reported. The author also performed impact analyses of the inputs to the model and reported that reservoir temperature has the greatest impact on the reservoir fluid viscosity. This is followed by oil density, pressure and gas/oil ratio as second, third and fourth respectively.

Oloso et al. (2009) used both FN and SVR to predict entire μ_o and R_s curves, covering the two properties below and above the saturation pressure. Instead of carrying out the usual data point prediction, the models predicted the parameters of the equations used to fit the curves. The models were built and tested using data sets from the Middle East crude oil PVT reports. In testing the models, AAPREs of 8.5514% and 10.2012% were reported for predicting viscosity and GOR

curves respectively for the FN models, while AAPREs of 8.5969% and 9.0757% were reported for the SVR models in predicting viscosity and GOR curves respectively.

Khoukhi (2012) developed hybrid models of GA and ANN, and GA and ANFIS to predict P_b and B_{ob}. In his experimentation, he reported to have used the same data sets used by Al-Marhoun (1988), Al-Marhoun & Osman (2002) and Osman & Al-Marhoun (2005). The authors indicated that the two hybrid models gave better performances than the compared regression correlations.

The authors in (Asoodeh & Bagheripour, 2012) developed a hybrid systems called: power-law committee with intelligent systems(PLCIS). The system integrates GA with the outputs of ANN, fuzzy logic (FL) and neuro-fuzzy (NF) to optimise the prediction of P_b. They developed and tested the models using 361 data points gathered from the following literature: (Ostermann & Owolabi, 1983; Dokla & Osman, 1992) Omar & Todd, 1993; De Ghetto & Villa, 1994). In testing the models, the reported CC of 0.975 for the PLCIS hybrid model. The CCs for the three standalone techniques: ANN, FL and NF were above 0.97 but lesser than 0.975. All the four techniques performed better than the compared empirical correlations.

Talebi et al. (2014) used multi-layer perceptron (MLP) ANN and RBF network to develop two different models to predict P_b. The gathered 750 PVT data points from the literature to train and test their models. They reported correlations coefficients of 0.94 and 0.95 for the MLP ANN and RBF network models respectively. They claimed that their models performed better than all the evaluated empirical correlations using the same data sets.

Considering the ML or computational techniques that have been used for prediction of PVT properties till date, none of the existing methods have utilised ensemble systems. One of the aims of the ensemble systems is to address the possible problem of non-global generalisation of an ML technique within its learning space. Also, the existing ML or computational techniques do not explore the relevance of API grouping of oils in building its learning model while this has been used to increase the accuracy of some empirical models.

When well designed, the effects of bias and variance in an ensemble systems are neutralised within the constituting base models. The easiest way to establish the stability of the model with respect to bias and variance is to test on multiple data sets.

This thesis aims to explore and investigate the capability of ensemble systems for prediction of PVT properties. Also, the impact of the API grouping of oils on prediction is investigated by performing intelligent grouping of the data sets before performing the actual ML learning and prediction on the input data sets.

Chapter 4. Research Methodology

4.1. Introduction

One of the main problems of standalone ML technique is the local minima. Solution for many of the ML techniques involves searching within input space that may become stuck in a local minimum which causes the model to perform poorly when presented with new data (Dietterich, 2000). An ML model could be stuck in local minima as a result of inefficient learning parameters or imbalanced datasets. Ensemble learning systems is one of the ways to address this problem.

Ensemble systems based on SVR and RTs are developed. The adopted schematic framework for the ensemble system is shown in Figure 4.1. The base models of both the ensemble SVR and ensemble RT are selected by a simple rule which has been devised to utilise both AAPRE and RMSE. In statistical analysis, no priority can conclusively be given to either error measure as there is no consensus on which one is superior (Chai & Draxler, 2014).

Each model is assigned two different ranks based on AAPRE and RMSE. The ranks are assigned in ascending order of both AAPRE and RMSE. As a rule of thumb, a consistent and stable model is expected to have the same ranks for both AAPRE and RMSE. This criterion tagged "Tying Ranking" is exemplified in Table 4.1. For instance, if 5 base models are desired to be selected to form the ensemble model, then all the first 5 models of Case 1 in Table 4.1 will be selected. For case 2, only the first 4 models will be used to build the ensemble model. Otherwise, the other option to meet the criterion of 'tying ranking' if 5 base models are still desired to form the ensemble system will be to rerun the experimentation, making sure that more or some of the base models' learning parameters have been tweaked. In essence, "tying ranking" selection criterion demands that there is no mismatch between the AAPRE and RMSE ranks of the models to be used to build the ensemble system.

Table 4.1 Example of "Tying Ranking"

Case 1					Case 2			
Model Number	AAPRE Rank	RMSE Rank	Decision		Model Number	AAPRE Rank	RMSE Rank	Decision
Model_1	1	1	chosen		Model_1	1	1	chosen
Model_2	2	2	chosen		Model_2	2	2	chosen
Model_3	3	3	chosen		Model_3	3	3	chosen
Model_4	4	4	chosen		Model_4	4	4	chosen
Model_5	5	5	chosen		Model_5	5	6	not chosen
Model_6	6	7	not chosen		Model_6	6	7	not chosen
Model_7	7	6	not chosen		Model_7	7	5	not chosen

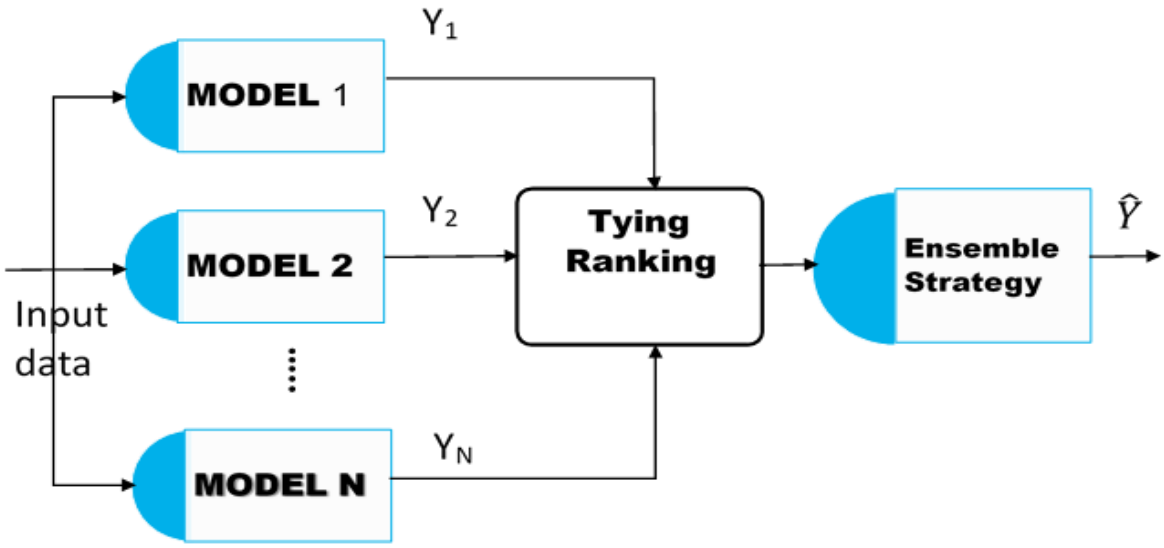


Figure 4.1. General Representation of the Implemented Ensemble Models

4.2. Ensemble Regression Trees and Support Vector Machine

The aim of this algorithm is to search within a large space of different inputs to dynamically determine the optimal parameters for both the RT and the ensemble average. *X* is the input data set and *Y* is the output vector. Both *X* and *Y* must have the same number of rows. *N* is the maximum number of branch nodes for each decision tree while *m* is the desired maximum number of trees in an ensemble model.

The implementation begins by randomly selecting *x* data set from the entire *X* data. *K* is set as the number of splits or branches in each RT. Different error tolerances (*j*'s) are experimented in building each RT. T_k creates a RT with maximum *k* splits. Each RT model is evaluated using AAPRE and RMSE. All the generated RT models are ranked separately based on AAPRE and RMSE in ascending orders. The "Tying ranking" criterion is used to select the base models which form the final ensemble RT.

If the number of base models desired to be in the ensemble does not meet the "tying ranking" criteria then either the number of base models is reduced or the experimentation is rerun. It will be observed that each experimental run does not use the same number of input data set. Stratification is performed where only 70% of the total input data points are used to train base models in each iteration. This allows different uncertainties to be possibly captured by different models.

1. Select randomly *x* data sets as 70% of the entire data sets (*X)* and the corresponding y from the output (*Y*)

2. Do for *k=1* to *N,* number of possible splits (e.g. *N=100*)

3. Do for *j*=1 to *m*, (set of error tolerance e.g. $10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}$)

4. Initialize a regression tree template,T_k

5. Compute RMSE and AAPRE for each tree

6. Continue until *j=m*

7. Continue until *k=N*

8. Choose the best *n* models for the ensemble using "tying ranking"

9. Predict the testing target *Y* from the testing input *x* using the *n* base RT models

10. Compute the ensemble output $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i$, where $\hat{Y}_i$ is the predicted target by the *i*th RT base models

Figure 4.2. Procedure for the Ensemble Regression Trees

Similar to the ensemble RT algorithm, the process of constructing ensemble SVR model in this work is shown in Figure 4.3. *X* and *Y* in Figure 4.3 are the inputs and the output respectively. *C* is the penalty factor which should neither be too large nor too small to prevent over-fitting and under-fitting respectively (Alpaydin, 2014). However, there is no definite criterion for determining its value as it is problem dependent. The value of ε controls the width of the ε-insensitive zone of the SVR model and determines its accuracy. The three main parameters that control each SVR model are $C,\ k$ and ε. Each SVR model $F(C, k, \varepsilon)$ is evaluated using AAPRE and RMSE. It should be noted that λ is another tuning parameter for the SVR model. However, based on preliminary investigations, λ was set to the same value as ε in each tuning iteration. To create the final ensemble SVR model, n number of the optimised base models will be selected. Before selecting the n base SVR models, ranks are assigned to all the base SVR models based on both AAPRE and RMSE in ascending orders. Only the base models which meet the "tying ranking" criterion are selected, otherwise, the procedure can be run with new dynamic parameters. Usually, the first 5 to 8 based models meet the criteria in this study. The output of the ensemble SVR will be the average of the outputs of the optimized SVR base models.

Meanwhile, if the same ranks for both *AAPRE* and *RMSE* cannot be established for the desired number of base models after several trials then the "Tying ranking" is constrained and the option of weighting ranking should be explored.

1. Select randomly *x* data sets as 70% of the entire data sets (*X)* and the corresponding y from the output (*Y*)
2. Iterate for *C=1* to *N (e.g. 1000)*
3. Iterate for kernel, $k \rightarrow \{RBF, Gaussian\}$
4. Iterate for $\varepsilon \in \{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$
5. Compute each SVR model $F(C, k, \varepsilon)$
6. Evaluate each SVR model using AAPRE and RMSE
7. Continue for the next ε
8. Continue for the next k
9. Continue until *C= N*
10. Choose the best *n* models for the ensemble using "tying ranking"
11. Predict the testing target *Y* from the testing input *X* using the *n* base SVR models
12. Compute the ensemble output $\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$, where $\hat{Y}_i$ is the predicted target by the *i*th SVR base model

Figure 4.3. Ensemble Support Vector Regression

4.3. Hybrid Functional Networks

Hybrid ML system is normally developed so that the consisting sub-systems complement each other. The aim of a hybrid system is to fill the gap that a single method cannot necessarily fill. This work implements a hybrid of K-means clustering and FN (Oloso et. al., 2017).

The K-Means Clustering algorithm is a simple and effective approach for producing clusters from a given data set. In essence, it finds the groups in a data set (Alpaydin, 2014) and the number of groups or clusters are predefined. Among other things, it is used for data pre-processing before performing classification or regression on a given data set. K-Means algorithm is given in Figure 4.4.

The outputs of the clusters are passed to the FNs. Each cluster is assigned a FN. The overall predictions from the different FNs are combined for final evaluation with indications or indices from the original data set before clustering which are then used for error evaluation between the predicted and the target values. The overall K-Means+FN hybrid system is depicted in Figure 4.5.

For $k = 1\ to\ K$
 do $\gamma_k \leftarrow initialise\ randome\ location$
end for
Repeat
 For $m = 1\ to\ M$
 $C_m \leftarrow argmin_k\ ||\gamma_k - x_m||$ $//x_m$ is a data point for given X data set
 end for
 For $k = 1\ to\ K$
 $X_k \leftarrow \{x_n : C_m = k\}$
 $\gamma_k \leftarrow mean(X_k)$
 end for
until γ converfe
return C // clusters

Figure 4.4. K-Means Clustering

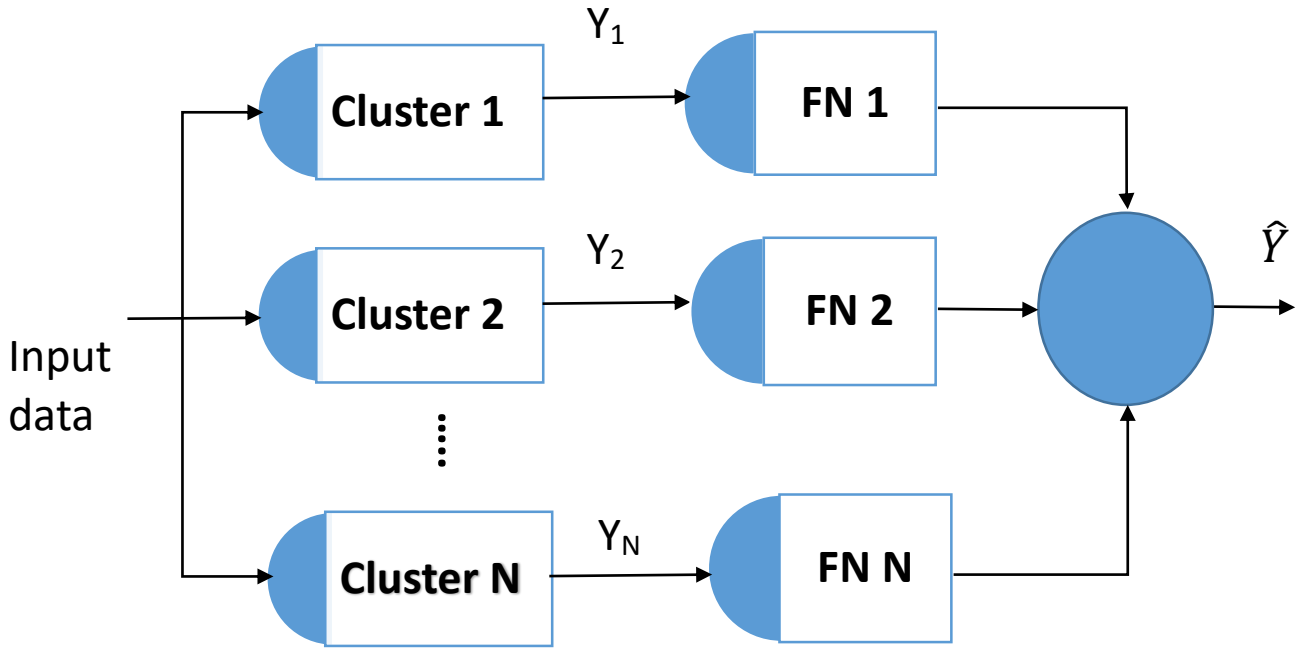


Figure 4.5. Overview of K-Means+FN

For all the PVT properties considered in this work, 4 clusters have been used. FN is then developed for each cluster as described in section 3.5. Whenever new data set is presented, the saved clustering algorithm puts them into the respective cluster which will then be handled by the respective FN for each cluster. The outputs from the FNs will then be combined to give the final and predicted output $\hat{Y}$.

The relationship for the *i*th cluster predictor can represented as follows.

$$Y_i = f(X_1, X_2, X_3, .., X_n) \quad (4.1)$$

The functional equation can then be given as:

$$y_i = \sum_{j=1}^{p} \sum_{r=1}^{m} C_r \varphi_r(x_{ij}), \quad i = 1,2,\dots,n \tag{4.2}$$

The linear independent function, φ_r, which has given stable and best results in all the consider PVT properties is a polynomial function of degrees 3 to 7.

After solving equation 4.2 using least squared optimisation method as highlighted in section 3.5., the final output of the K-Means+FN system is the union of all the cluster predictors. The process for developing K-Means+FN is summarised in

1. Input: Get the data *X* for clustering
2. Data clustering with K-Means to get clusters, D_c, *c*=1,..,*k* where *k* is total number of clusters
3. BEGIN FN Network
3. FOR *m*=3 to 10
4. Initialise the functional model e.g. $\varphi = \{1, x, x^2, \dots, x^m\}$
5. FOR each cluster D_c compute the functional components *C*
6. END FOR *m* loop
7. END each cluster loop
8. Combine and evaluate the prediction from the clusters.

Figure 4.6. Implementation of hybrid K-Means+FN

4.4. Experimental Data Sets

There are 11 different data sets used in this study. They are labelled A to I and their statistical descriptions are shown in the appendix. Some of these data had been acquired from the literature. The first dataset A has 895 data points with 327 data points from different published works (Al-Marhoun, 1988; Dokla & Osman, 1992; Omar & Todd, 1993; Bello et. al., 2008), while the remaining part of data set A is from unpublished PVT data which includes data from GeoMark Research Limited. Data set B has 1200 data points which includes data set A, independent sources and PVT reports from Shell Petroleum Development Company, Nigeria. The remaining data sets (C-K) are mainly from GeoMark Research Limited, some independent sources and Shell Petroleum Development Company, Nigeria.

Chapter 5. Results and Discussions

5.1. Introduction

Different experimentations have been set up to build six ML models for the prediction of P_b, B_{ob}, μ_{od}, μ_{ob} and μ_{oa}. The developed ML models are described in sections 4.2 and 4.3. For P_b and B_{ob}, two different models have been developed using different data sets. The models from these two experiments are then applied to three different data sets for testing of generalization, reliability and sensitivity of the models to different data sets. Almost all the empirical correlations for these two PVT properties use the same functional form, i.e. the same independent variables, that was proposed by Standing (1947). Hence, only this functional form has been investigated.

On the other hand, different functional forms that can be used to formulate μ_{od}, μ_{ob} and μ_{oa} have been investigated by building different models for each of the functional form with the six ML techniques. Two different data sets are subsequently used to test the models.

Four common statistical measures (CC, $RMSE$, E_a, and E_{max}) have been reported for in the prediction results of all the models. SD has not reported because of the space for graphical visualisation and it has not been used for result evaluation because it only shows the distribution of the errors but it does not necessarily indicate how low the errors are relative to other set of errors. Also, very low E_{min} may potentially mean an over-fitting though its low value is desired, it is not an adequate condition for the performance of a model to be good. More information will be given on this later.

5.2. Prediction of Bubblepoint Pressure

Two different experiments had been carried out which involved developing different ML techniques using data sets A and B respectively. The functional form that was used in all developing ML models for this property is $(R_s, \gamma_g, \gamma_{API}, T)$. For each of the models described in chapter 4 and their standalone algorithms, two different models had been developed to test the capability and influence of training the ML techniques with input data sets of different sizes and variations. These two models for each ML technique were then tested on additional three different data sets (C, D and E) to investigate the stability and generalisation capability of the models. The selected empirical correlations were also tested on these data sets.

5.2.1. Results and Error Analysis for Bubblepoint Pressure

Results for prediction of P_b using selected empirical correlations and the six ML techniques are presented in Figures 5.1 to 5.8. Common statistical measures for PVT analysis are presented for

each experiment or evaluation category. Figures 5.1 and 5.2 show the results of experiments 1 and 2 when different data sets were used to train the ML techniques.

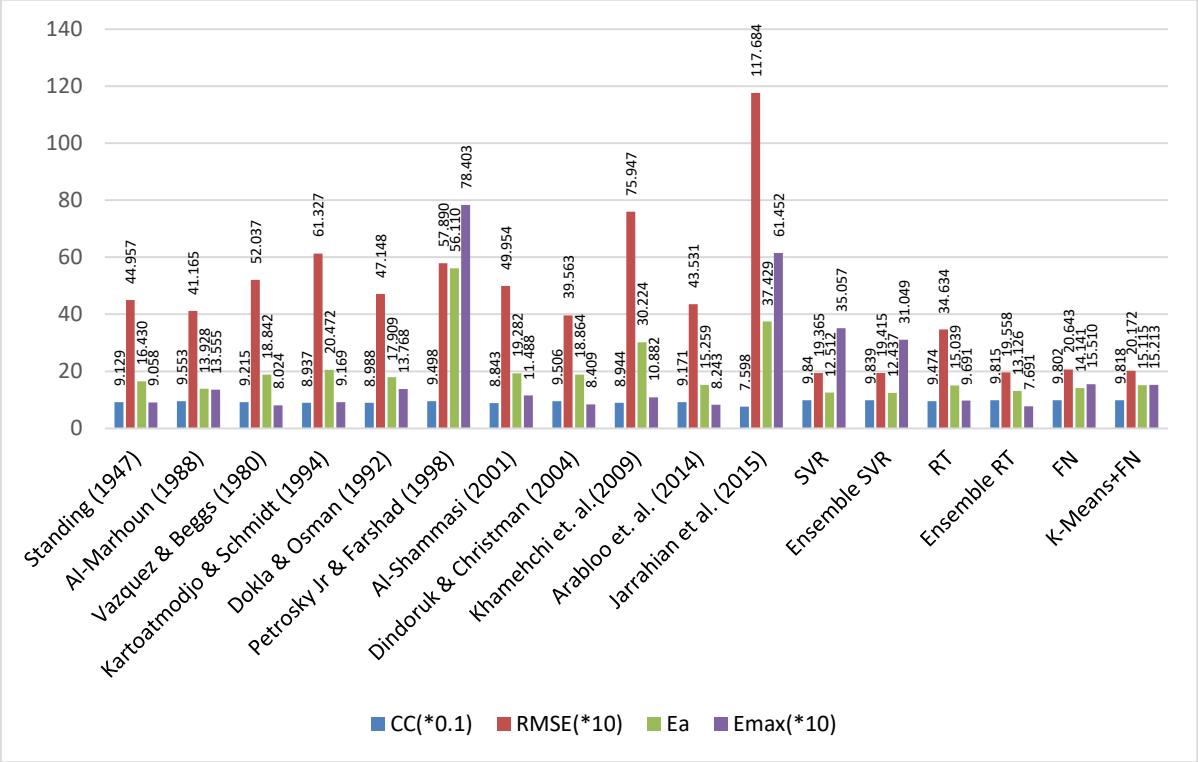


Figure 5.1. Performances of empirical correlations and ML techniques for P_b prediction using data set A (Experiment 1)

In Figure 5.1, standalone SVR has the least RMSE (193.6493) though graphically, the *RMSE* of all the ML techniques are very similar. The ensemble SVR has the least E_a. The overall performance of standalone SVR and ensemble SVR are very competitive. On the other hand, the ensemble RT has a significantly higher performance than the standalone RT with very conspicuous lower *RMSE* . Also, the hybrid K-Means+FN shows a slightly better performance than FN with slightly larger CC, smaller RMSE and E_{max}, though FN has smaller E_a than its hybrid model. In this comparative stage, standalone SVR, ensemble SVR, ensemble RT, FN and hybrid K-Means+FN outperform all the compared empirical correlations.

Meanwhile, in Figure 5.1, the E_{max} for both standalone SVR and ensemble SVR are large, even larger than many empirical correlations. Though this is undesirable, it does not necessarily mean overall bad results since their other statistical measures are good. In fact, it is probably that the error of a single data point gave E_{max} that is very large while others are small. This is evident in their overall smaller average error (E_a) than many empirical correlations. Similar explanation may

hold for both FN and K-Means whose E_{max} are considerably larger than the E_{max} for some empirical correlations though their other statistical measures are better than many of them. In contrast, the ensemble RT model gives the least E_{max} (76.9121) among all the methods in Figure 5.1.

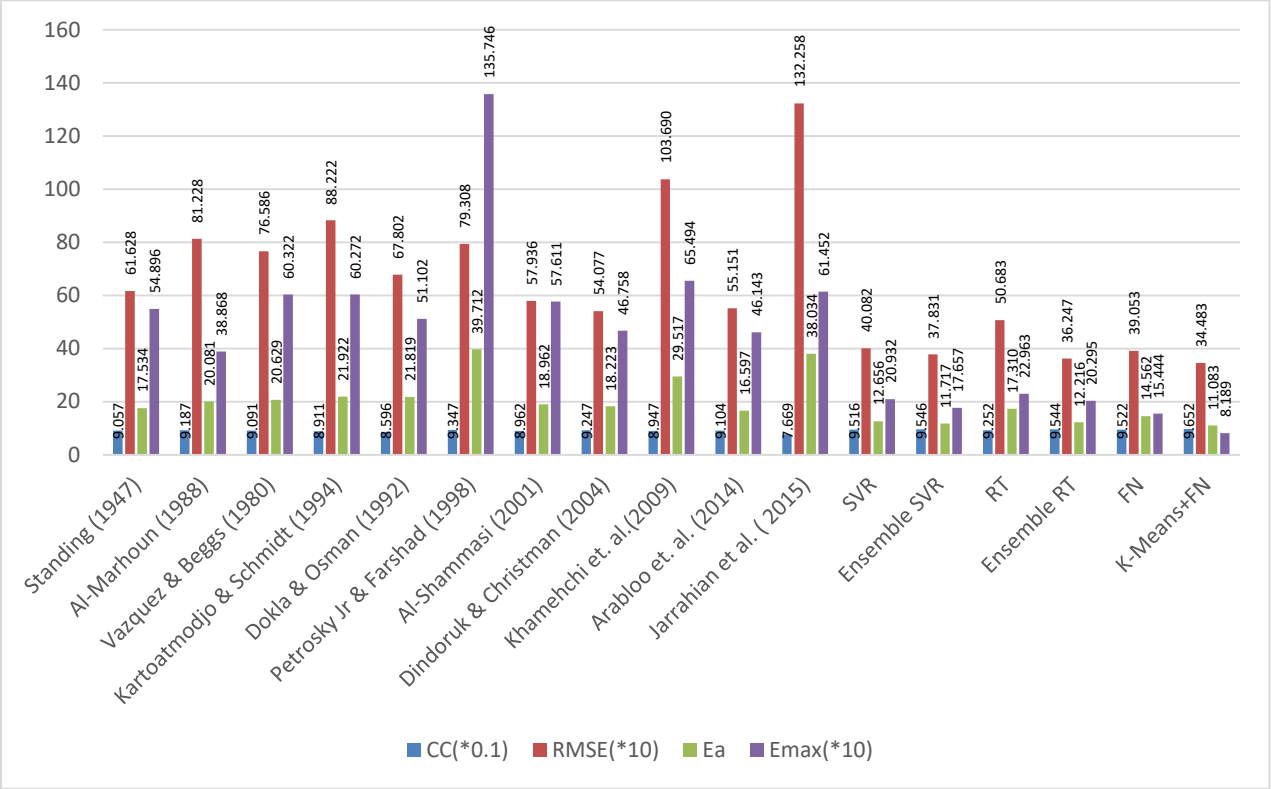


Figure 5.2. Performances of empirical correlations and ML techniques for P_b prediction using data set B (Experiment 2)

The validation results of experiment 2 where a larger and more diverse data set has been used to train the six ML techniques are shown in Figure 5.2. The overall best performance is given by hybrid K-Means+FN with the largest CC, smallest RMSE, E_a and E_{max}. The second best performance is between ensemble RT and ensemble SVR with competitive results. Ensemble SVR has the second largest CC and the second smallest E_a while ensemble RT has the second smallest RMSE. Ensemble RT shows a higher performance than the standalone RT model. Five ML models – standalone SVR, ensemble SVR, ensemble RT, FN and hybrid K-Means+FN have better performance than all the compared empirical correlations. The *RMSE* of empirical correlation by Jarrahian et al. (2015) and the E_{max} by Petrosky Jr & Farshad (1998) are typically high among all the methods for this data set.

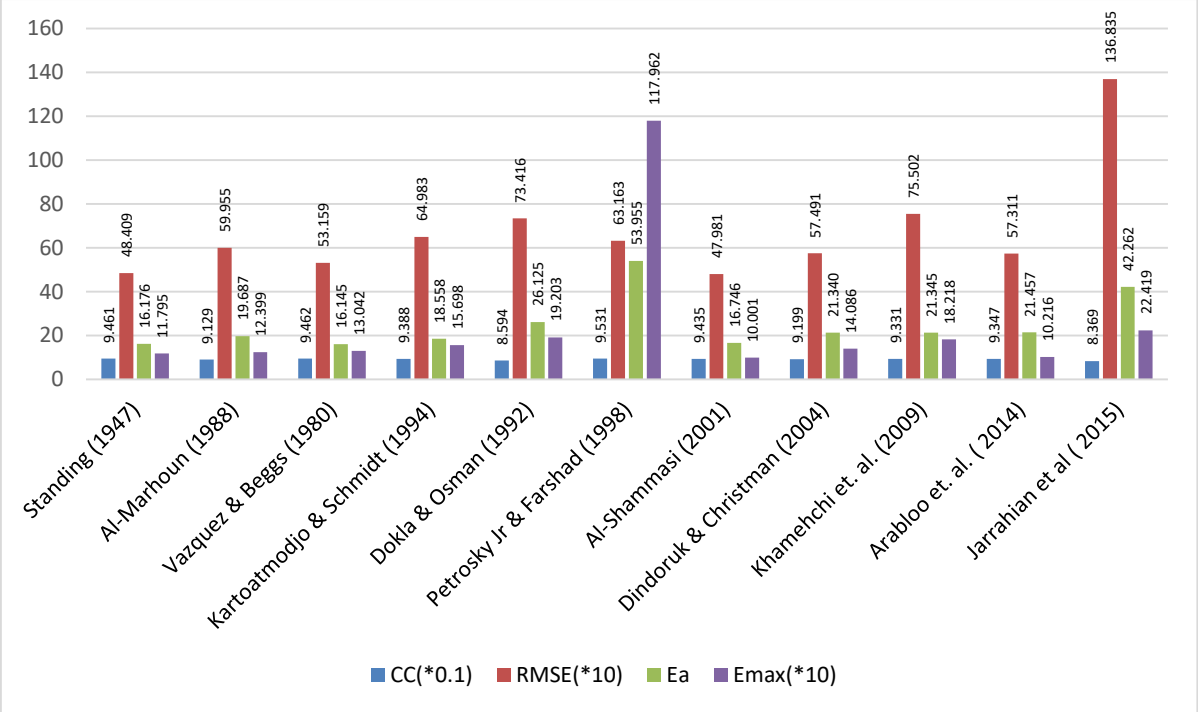


Figure 5.3. Performances of empirical correlations for P_b prediction using data set C

Figure 5.3 shows the performances of the selected empirical correlations for P_b prediction on data set C. Among all the correlations, Standing (1947), Vazquez & Beggs (1980) and Al-Shammasi (2001) have competitive and leading results across the evaluating four statistical measures. It can be observed that though Petrosky Jr & Farshad (1998) has the largest CC, however, its error measures (RMSE, E_a and E_{max}) are larger than the corresponding values for many of the correlations in Figure 5.3 especially the first best three. This type of behaviour will be analysed further in the next section.

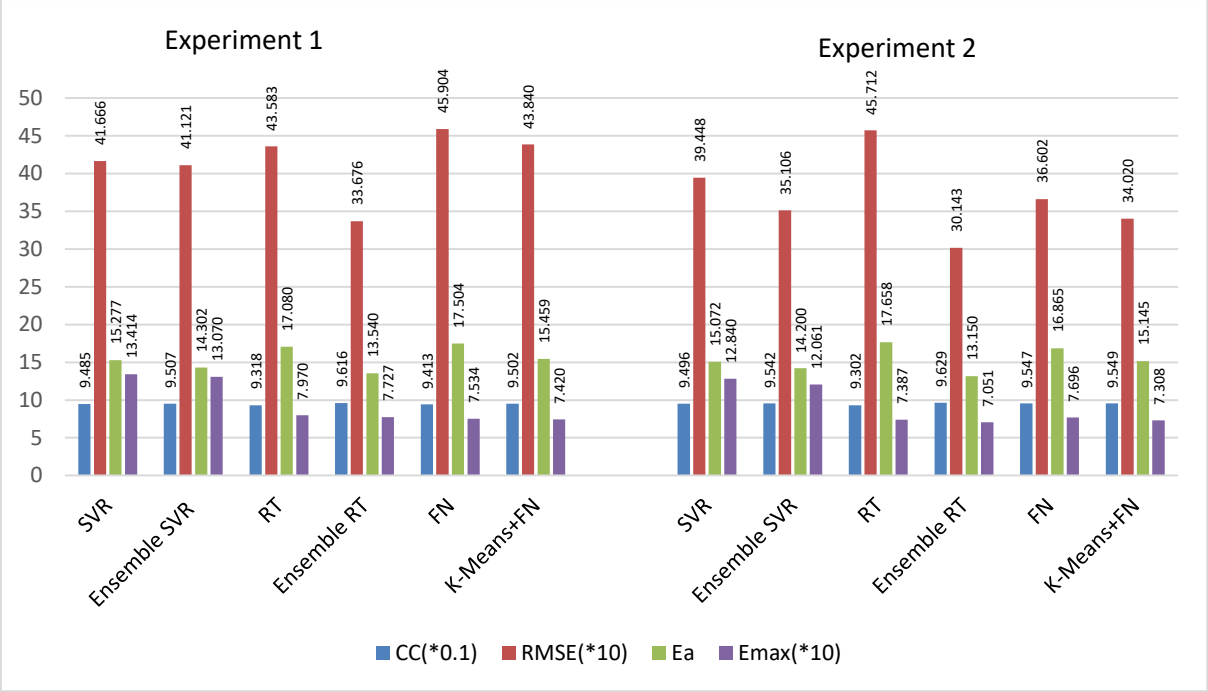


Figure 5.4. Performa2nces of ML techniques (Experiments 1 and 2) for P_b prediction using data set C

Figure 5.4 shows the testing results for all the ML models on dataset C. The ensemble RTs from both experiments 1 and 2 show significant performance improvements over their corresponding standalone RT models. The ensemble RT in experiment 2 has the largest CC, lowest RMSE, E_a and E_{max}. This is followed by the performance of K-Means+FN from experiment 2.

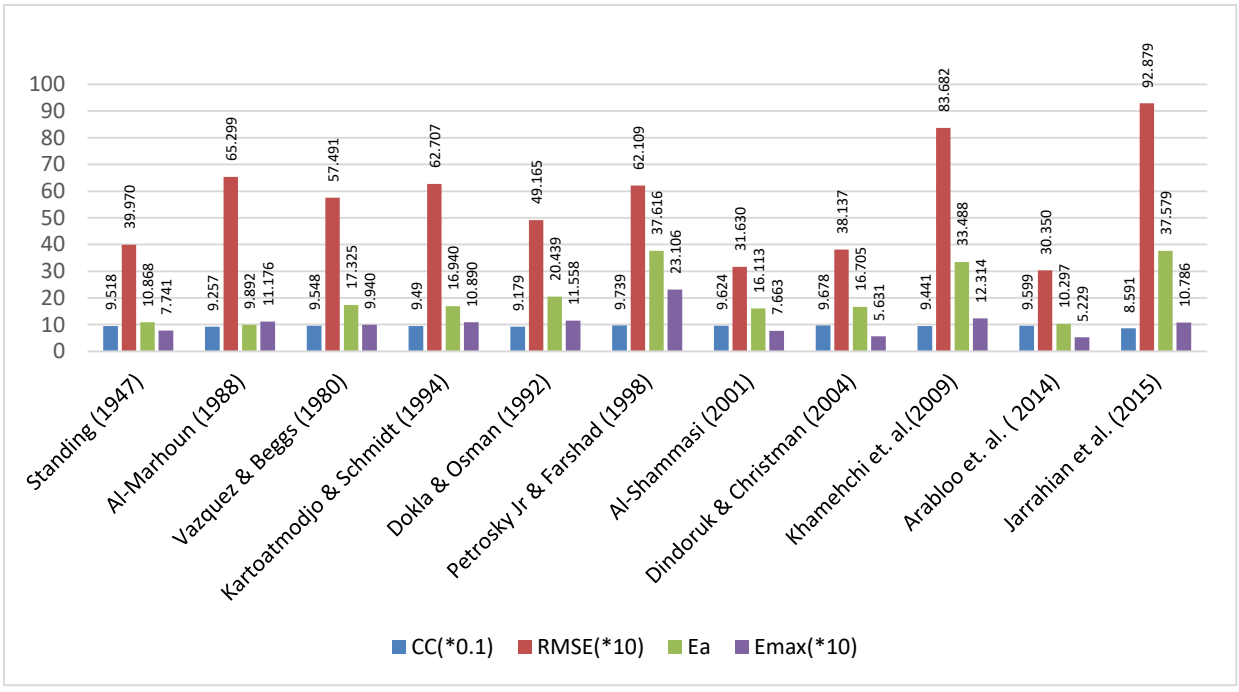


Figure 5.5. Performances of empirical correlations for P_b prediction using data set D

Figure 5.5 presents the performances of empirical correlations for P_b prediction on data set D. Among the correlations, Arabloo et. al. (2014) has the best performance with the smallest RMSE, E_a and E_{max} though its CC is not the largest but it is competitive. The correlation of Petrosky Jr & Farshad (1998) which gives the largest CC has not shown very competitive results across other evaluation parameters. The worst performance is given by Jarrahian et al. (2015) with high RMSE, E_a and E_{max} compared to others.

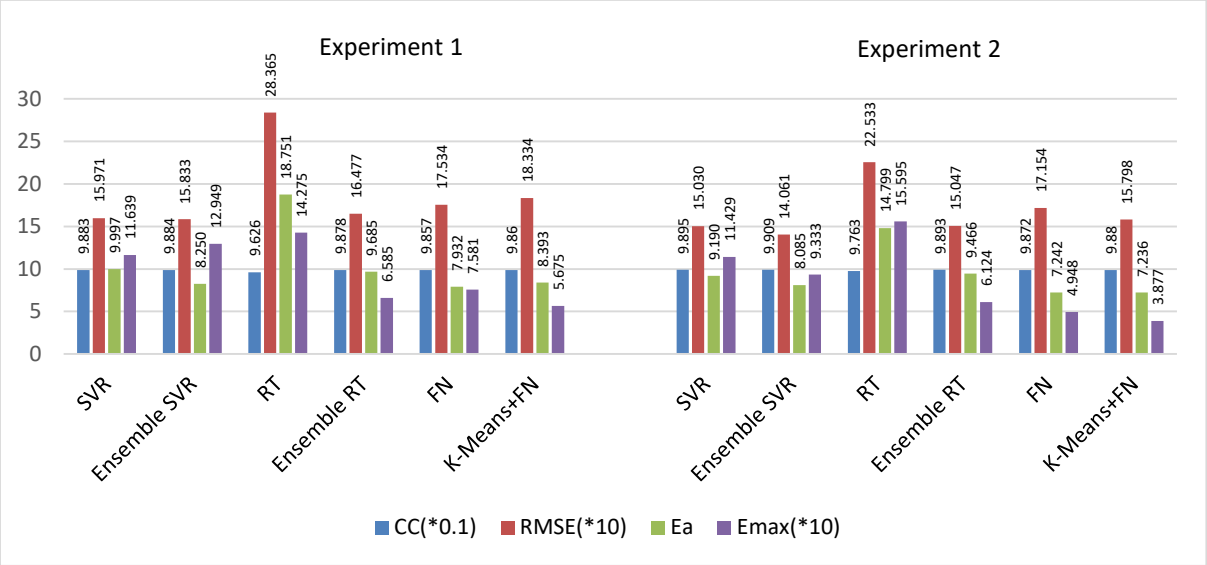


Figure 5.6. Performances of ML techniques (Experiments 1 and 2) for P_b prediction using data set D

The comparative results of P_b prediction from dataset D for ML techniques are shown in Figure 5.6. The overall results for all models are generally better than dataset C. The ensemble SVR from experiment 2 has the largest CC with competitive CC values from standalone SVR and ensemble RTs from both experiments, and K-Means+FN from experiment 2. The ensemble SVR from experiment 2 also has the lowest *RMSE* while K-Means+FN from experiment 2 has the lowest E_a and E_{max}. The overall best performance on this data set lies between ensemble SVR and K-Means+FN from experiment 2.

Comparing with the empirical correlations, Arabloo et. al. (2014), which is the best performing model among the empirical correlations, gives E_{max} that is lower than the corresponding values for all ML techniques except K-Means+FN on this data set, though its overall performance is still lower than the results of ensemble SVR, standalone SVR, ensemble RT, FN and K-Means+FN given by its lower CC, higher RMSE and E_a.

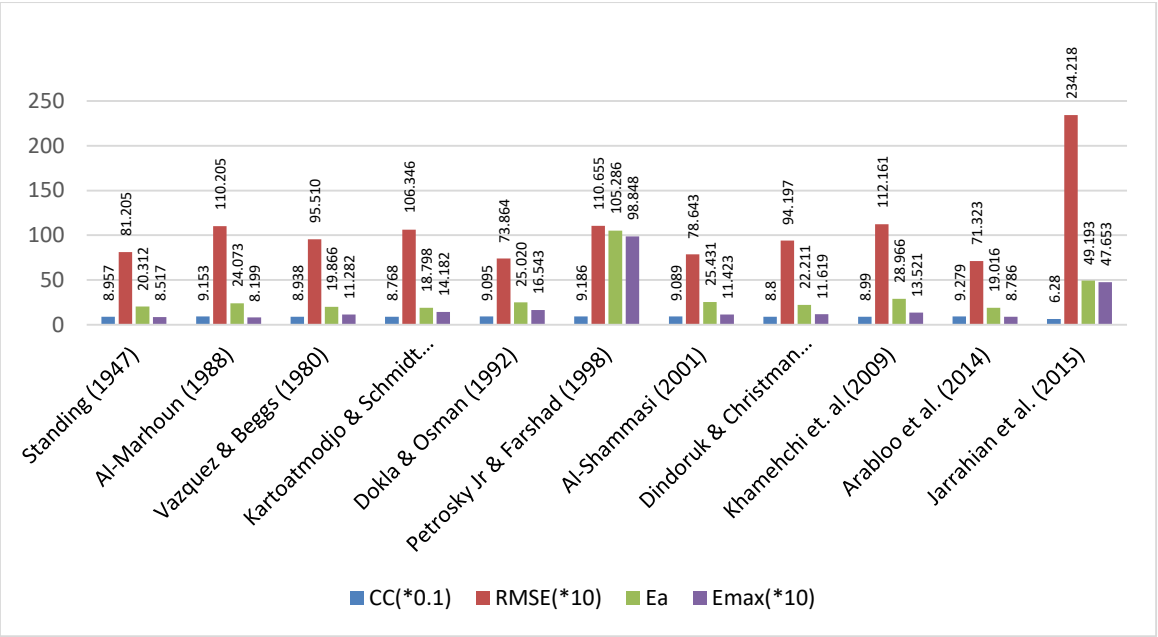


Figure 5.7. Performances of empirical correlations for P_b prediction using data set E

Figure 5.7 presents the performances of empirical correlations for P_b prediction on data set E. Arabloo et al. (2014) gives the maximum CC and minimum RMSE with the second lowest E_a among the empirical correlations. Jarrahian et al. (2015) gives a notably large RMSE while Petrosky Jr & Farshad (1998) gives a very high E_{max}, indicating high offset in the prediction of some data points. The performances of the empirical correlations are generally poor on this data set.

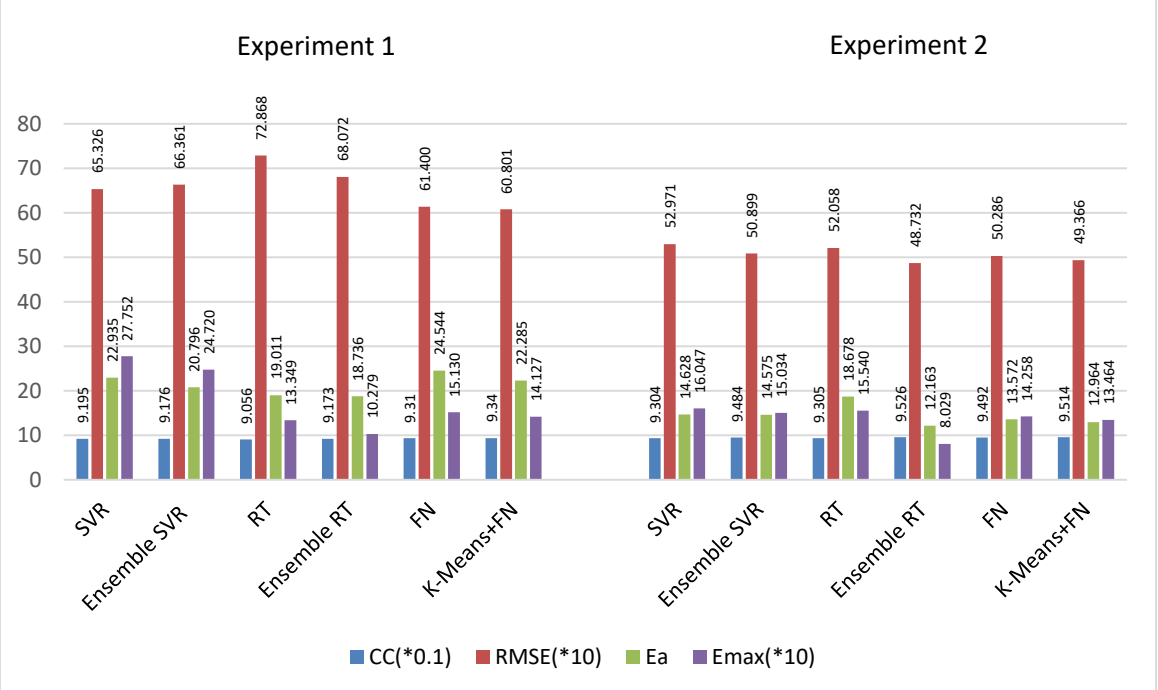


Figure 5.8. Performances of ML techniques (Experiments 1 and 2) for P_b prediction using data set E

Figure 5.8 shows the results of the ML models for the prediction of P_b using dataset E. The ensemble RT of experiment 2 has the highest CC and minimum statistical errors: RMSE, E_a and E_{max}. The ensemble RT in experiment 2 shows a very high improvement across all statistical parameters compared to the standalone RT and ensemble RT in experiment 1. The results for ensemble SVR and standalone SVR in experiment 2 also increased significantly compared to their corresponding methods in experiment 1. K-Means+FN also has better performance than the standalone FN.

Comparing the ML models with the empirical correlations for data set E, Al-Marhoun (1988) gives a very competitive E_{max} which is lower than corresponding value for some ML techniques, though its other error measures, RMSE and E_a, are much higher than the corresponding values for many ML models. Some other empirical correlations (Standing, 1947; Vazquez & Beggs, 1980; Kartoatmodjo & Schmidt, 1994; Arabloo et al., 2014) also show considerably low E_a but their other performance measures are also lower than many of the ML models in the two experiments as they have lower CC and/or larger RMSE/E_{max}.

5.2.2. Trend and Comparative Analysis of Bubblepoint Pressure Prediction

To get an overview of the overall performances of all the developed ML models and the comparing empirical correlations, a summary of the minimum and maximum values of each of the statistical measures (CC, $RMSE$, E_a and E_{max}) used in the performance analysis is given In Tables 5.1 and 5.2 for the empirical correlations and the ML models respectively. It can be recalled that two models have been built for each ML technique under consideration of P_b using 70% of data sets A and B respectively. Consequence upon this, to allow fair comparisons, only results of data sets C, D and E are considered in Tables 5.1 and 5.2.

It can also be recalled that the higher the value of CC the better is the result. On the other hand, the lower the values of ($RMSE$, E_a and E_{max}), the better is the result. The range of the statistical measures can also be useful in detecting or checking the consistency of a method in relation to others, though that statistical measure, i.e. range, cannot solely determine the ranking of a method's performance. This is because a poor performing method may have a short range for a particular statistical measure and still has a poor performance than other comparing methods.

Table 5.1. Summary of Statistical Measures for P_b Empirical Correlations

Correlation Method	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
Standing (1947)	0.8957/ 0.9518	399.70 / 812.05	10.87 / 20.31	77.41 / 117.95
Al-Marhoun (1988)	0.9129/ 0.9257	599.55 / 1102.05	9.89 / 24.07	81.99 / 123.99
Vazquez & Beggs (1980)	0.8938/ 0.9548	531.59 / 955.10	16.14 / 19.87	99.40 / 130.42
Kartoatmodjo & Schmidt (1994)	0.8768/ 0.9490	627.07 / 1063.46	16.94 / 18.80	108.90 / 156.98
Dokla & Osman (1992)	0.8594/ 0.9179	491.65 / 738.64	20.44 / 26.12	115.58 / 192.03
Petrosky Jr & Farshad (1998)	0.9186/ 0.9739	621.09 / 1106.55	37.62 / 105.29	231.06 / 1179.62
Al-Shammasi (2001)	0.9089/ 0.9624	316.30 / 786.43	16.11 / 25.43	76.63 / 114.23
Dindoruk & Christman (2004)	0.8800/ 0.9678	381.37 / 941.97	16.71 / 22.21	56.31 / 140.86
Khamehchi et. al.(2009)	0.8990/ 0.9441	755.02 / 1121.61	21.34 / 33.49	123.14 / 182.18
Arabloo et al. (2014)	0.9279/ 0.9599	303.50 / 713.23	10.30 / 21.46	52.29 / 102.16
Jarrahian et al. (2015)	0.6280 / 0.8591	928.79 / 2342.18	37.58 / 49.19	107.86 / 476.53

Table 5.2. Summary of Statistical Measures for Testing P_b Prediction with ML techniques

Experiment 1	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
SVR	0.9195 / 0.9883	159.71 / 653.26	9.997 / 22.935	116.39 / 277.52
Ensemble SVR	0.9176 / 0.9884	158.33 / 663.61	8.250 / 20.796	129.49 / 247.20
RT	0.9056 / 0.9626	283.65 / 728.68	17.080 / 19.011	79.70 / 142.75
Ensemble RT	0.9173 / 0.9878	164.77 / 680.72	9.685 / 18.736	65.85 / 102.79
FN	0.9310 / 0.9857	175.34 / 614.00	7.932 / 24.544	75.34 / 151.30
K-Means+FN	0.9340 / 0.9860	183.34 / 608.01	8.393 / 22.285	56.75 / 141.27
Experiment 2	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
SVR	0.9304 / 0.9895	150.30 / 529.71	9.190 / 15.072	114.29 / 160.47
Ensemble SVR	0.9484 / 0.9909	140.61 / 508.99	8.085 / 14.575	93.33 / 150.34
RT	0.9302 / 0.9763	225.33 / 520.58	14.799 / 18.678	73.87 / 155.95
Ensemble RT	0.9526 / 0.9893	150.47 / 487.32	9.466 / 13.150	61.24 / 80.29
FN	0.9492 / 0.9872	171.54 / 502.86	7.242 / 16.865	49.48 / 142.58
K-Means+FN	0.9514 / 0.9880	157.98 / 493.66	7.236 / 15.145	38.77 / 134.64

Considering the summary reports in Tables 5.1 and 5.2, the following can be inferred.

- No empirical correlation has a minimum or maximum CC comparable with all the ML models in experiment 2.
- No empirical correlation has a maximum CC comparable with all the ML models in experiments 1 except the RT model. Regarding the maximum CC where Petrosky Jr & Farshad (1998) has higher maximum CC than the RT from experiment 1, it will be noticed that its error measures are comparably higher.
- With respect to both minimum and maximum RMSE, all ML models from both experiments 1 and 2 have lower respective RMSE values.
- Few empirical correlations (such as Standing (1947)) have competitive results for either or both minimum and maximum E_a, their CC and RMSE have put their overall performances at disadvantage. In fact, all ML models from experiment 2, except RT, still outperform this empirical correlation and all others since those ML models have smaller minimum and maximum E_a.
- E_{max} appears to be the main statistical measure for which the empirical correlations compete most fiercely with the ML models from both experiments. Only FN and K-Means have their minimum E_{max} smaller than the corresponding values of all other methods. However, it can clearly be observed that some of these empirical correlations have performed worse with respect to other statistical measures.
- Given all these aforementioned points, it can be concluded that most ML models from experiment 2 and some from experiment 1 have better generalisation capabilities and perform better than the comparing empirical correlations

Likewise, Table 5.3 compares the standalone ML techniques with their respective ensemble system or hybrid system. The overall best performing method on each data set for the bubblepoint pressure prediction has also been indicated. This summary report is based on the results presented in section 5.3.1. Recall that four main statistical measures have been considered in the result analysis, hence, the possibility of a tie between methods in a category or the overall results.

It can be observed that some empirical correlations have given competitive results in some cases with the ML techniques, notably data sets C and D as well data set A to an extent. The maximum γ_{API} for the data sets A, C, D are 55, 63.7 and 44.93 respectively which are within the limit stated in the literature (McCain, 1990) for black oil, or not too far from it. Considering data sets B and E, the maximum γ_{API} are 105.02 and 115 which are much higher than the typical maximum value of γ_{API} for black oils.

In relation to the ML techniques from experiments 1 and 2 which had been built on 70% of data sets A and B respectively, it will be observed that the models from the latter experiment generally perform better for data sets C, D and E in predicting P_b.

Table 5.3. Summary Comparison of ML Technique for Bubblepoint Pressure Prediction

	Data Sets				
Methods	A	B	C	D	E
SVR vs Ensemble SVR	Tie between SVR and Ensemble SVR	Ensemble SVR	Ensemble SVR	Ensemble SVR	Ensemble SVR
RT vs Ensemble RT	Ensemble RT	Ensemble RT	Ensemble RT	Ensemble RT	Ensemble RT
FN vs K-Means+FN	K-Means+FN	K-Means+FN	K-Means+FN	K-Means+FN	K-Means+FN

In corroboration to the summary reports of Tables 5.1 and 5.2 and the inferences from them, Figures 5.9 and 5.10 show the average statistical measures for P_b prediction from both empirical correlations and the six ML models respectively. Recall that higher value of CC implies possible better performance, while in contrary, lower values of the errors (CC, $RMSE$ and E_a) are desired and technically mean better accuracy.

At least five ML models from experiment 2 (SVR, ensemble SVR, ensemble RT, FN and K-Means+FN) clearly outperform all the empirical correlations with respect to the average CC, $RMSE$ and E_a. Though some empirical models have low average values of E_{max} which are competitive with the corresponding values of the ML models, their other statistical measures are clearly poorer.

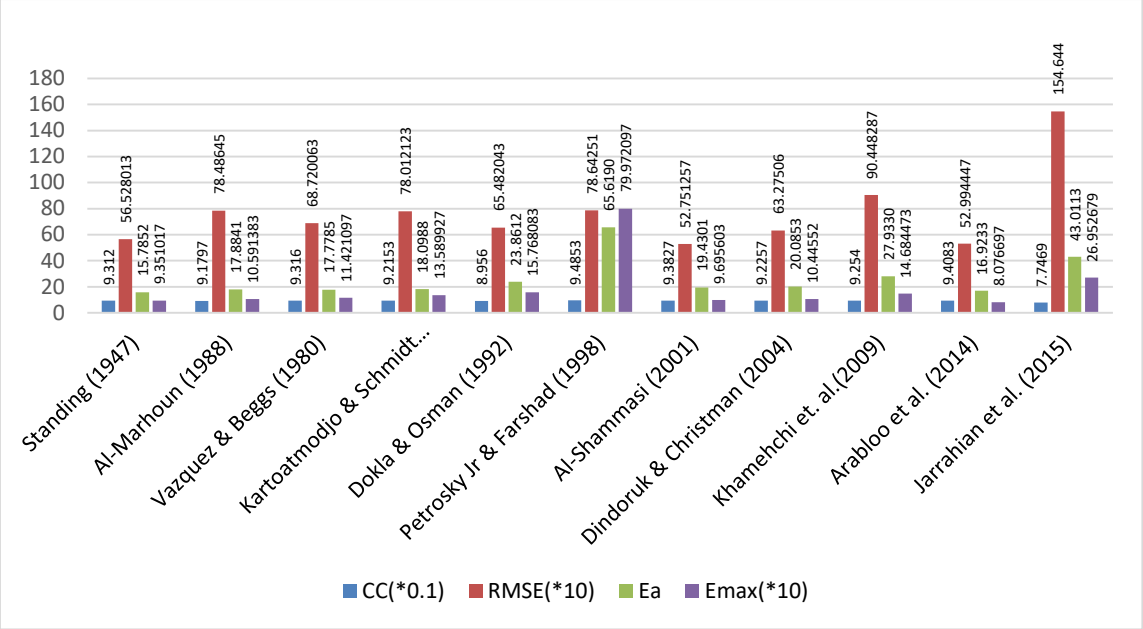


Figure 5.9. Average of Statistical Measures for P_b Empirical Correlations

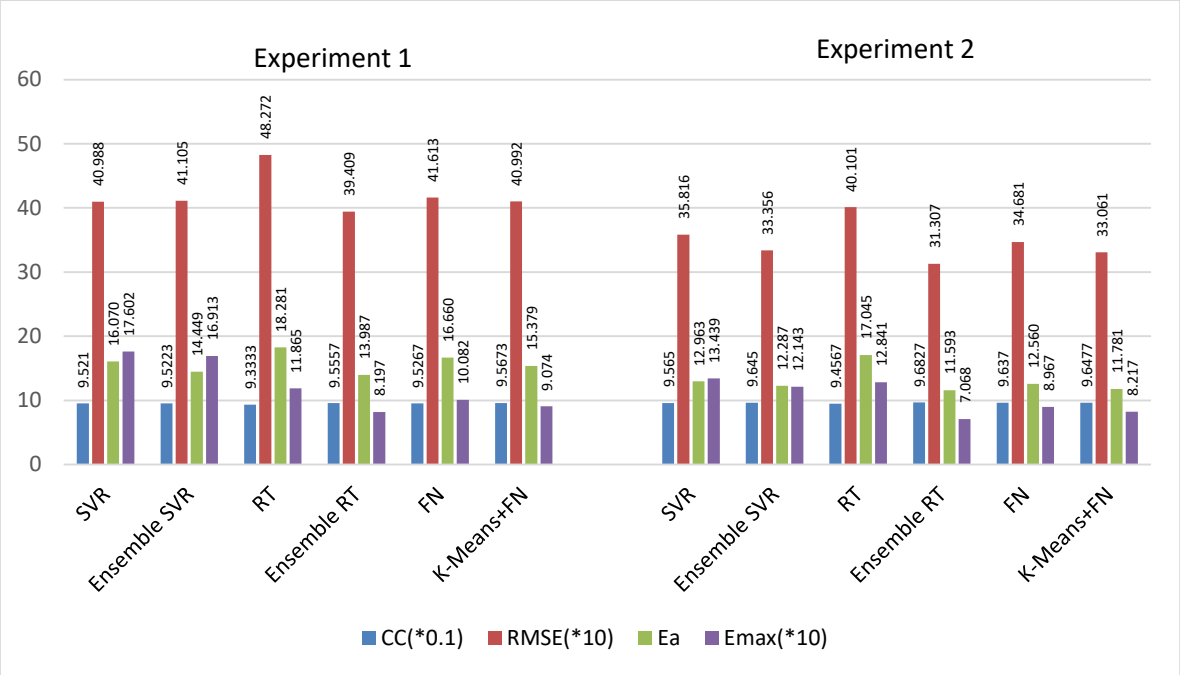


Figure 5.10. Average of Statistical Measures for P_b Prediction with ML techniques

5.3. Prediction of Oil Formation Volume Factor at Bubblepoint Pressure

Similar to the P_b experimental work, the functional form adopted for developing all the ML models for B_{ob} prediction is $f(R_s, \gamma_g, \gamma_{API}, T)$.

5.3.1. Results and Error Analysis for Oil FVF at Bubblepoint Pressure

The results for the prediction of B_{ob} using selected empirical correlations and the six ML techniques are presented in Figures 5.11 to 5.18. Similar to P_b analysis, Figures 5.11 and 5.12 show the results for both experiments 1 and 2 when the ML techniques were trained using 70% of data sets A and B respectively. The developed models from the two experiments are subsequently tested on data sets C-E.

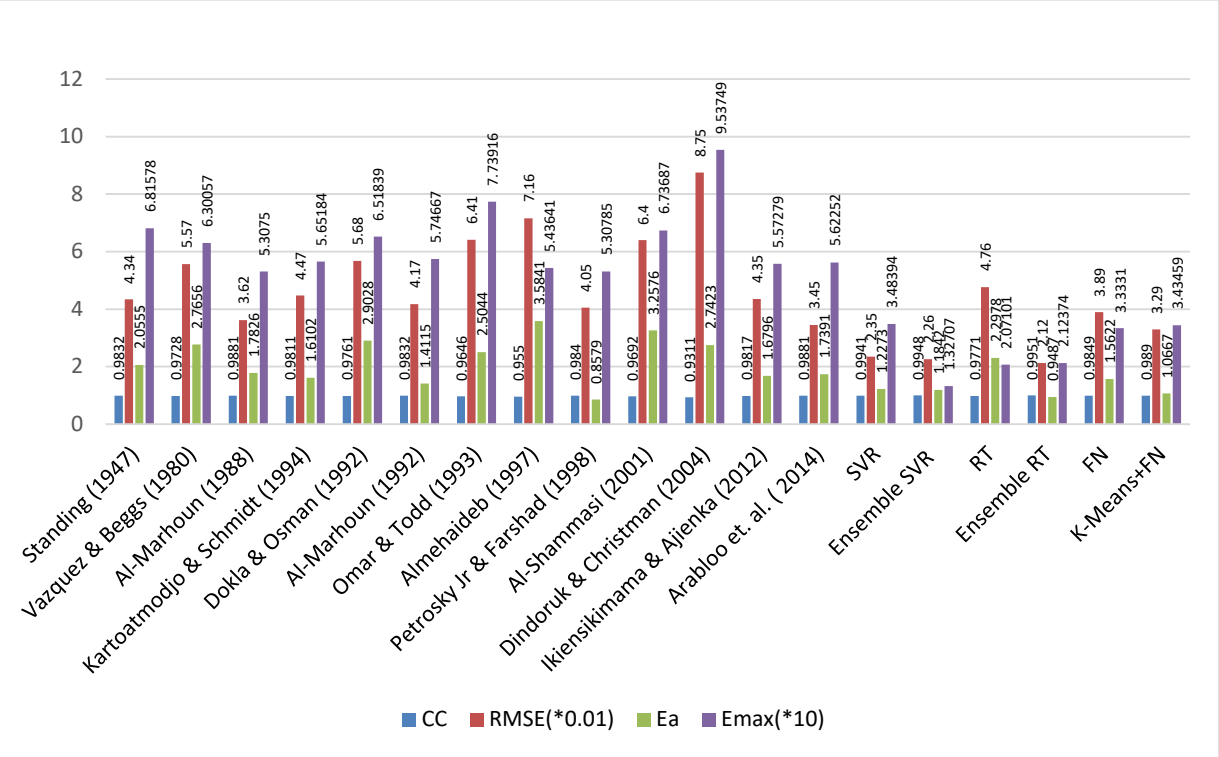


Figure 5.11. Performances of empirical correlations and ML techniques on data set A for B_{ob} Prediction (Experiment 1)

Figure 5.11 shows the results of selected empirical correlations and the six ML techniques for B_{ob} prediction in experiment 1. Ensemble RT has the highest CC and the lowest RMSE. Though the correlation of Petrosky Jr & Farshad (1998) gives the minimum E_a, its overall performance trails many ML techniques as they have better CC, $RMSE$ and E_{max}. The ensemble SVR has the smallest E_{max} but its overall performance is lower than ensemble RT. The performance of the standalone SVR slightly trails that of the ensemble SVR, while ensemble RT shows clearer and wider performance improvement over the standalone RT.

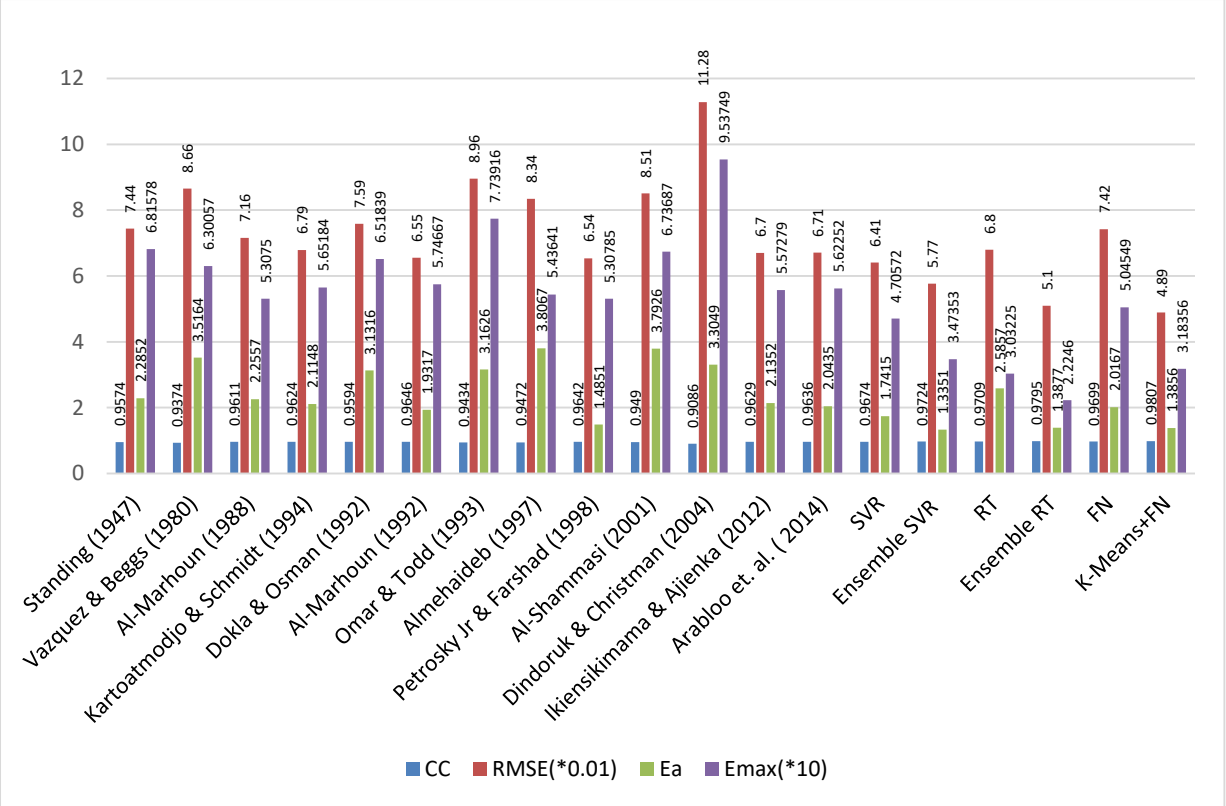


Figure 5.12. Performances of empirical correlations and ML techniques on data set B for B_{ob} Prediction (Experiment 2)

Figure 5.12 shows the results of selected empirical correlations and the six ML techniques for B_{ob} prediction in experiment 2. The hybrid K-Means+FN has the highest CC and lowest $RMSE$ with competitive E_a and E_{max}. The ensemble RT gives the lowest E_{max} (22.2460) among all the ML techniques and empirical correlations, and its $RMSE$ and E_a are very competitive with those of K-Means+FN. In this experiment, the ensemble SVR, ensemble RT and K-Means+FN – all perform better than their corresponding standalone techniques. The performances of all the models are averagely good and the poorest performance for this data set is given by Dindoruk & Christman (2004).

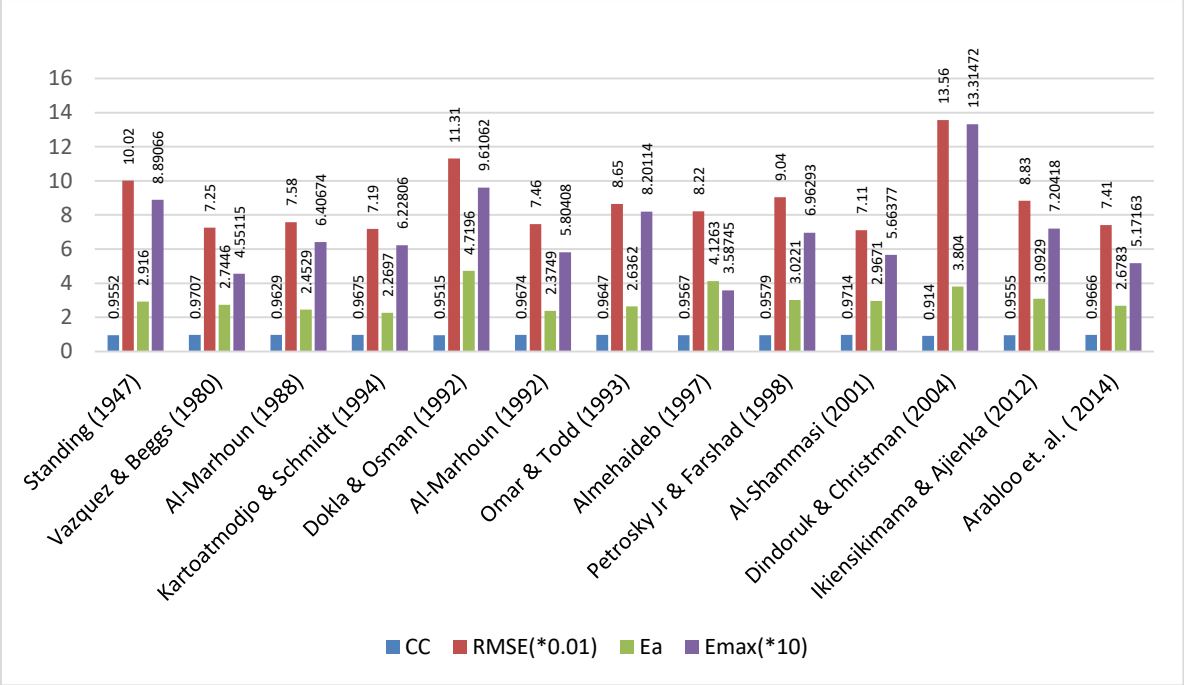


Figure 5.13. Performances of empirical correlations on data set C for B_{ob} Prediction

The performances of empirical correlations on data set C for B_{ob} Prediction are shown in Figure 5.13. On this data set, Al-Shammasi (2001) gives the largest CC and lowest RMSE while Kartoatmodjo & Schmidt (1994) gives the lowest E_a. Almehaideb (1997) gives lowest E_{max} though its other performance measures are not as good as the previous two correlations. Dindoruk & Christman (2004) gives the worst performance with the smallest CC, largest RMSE and E_{max}.

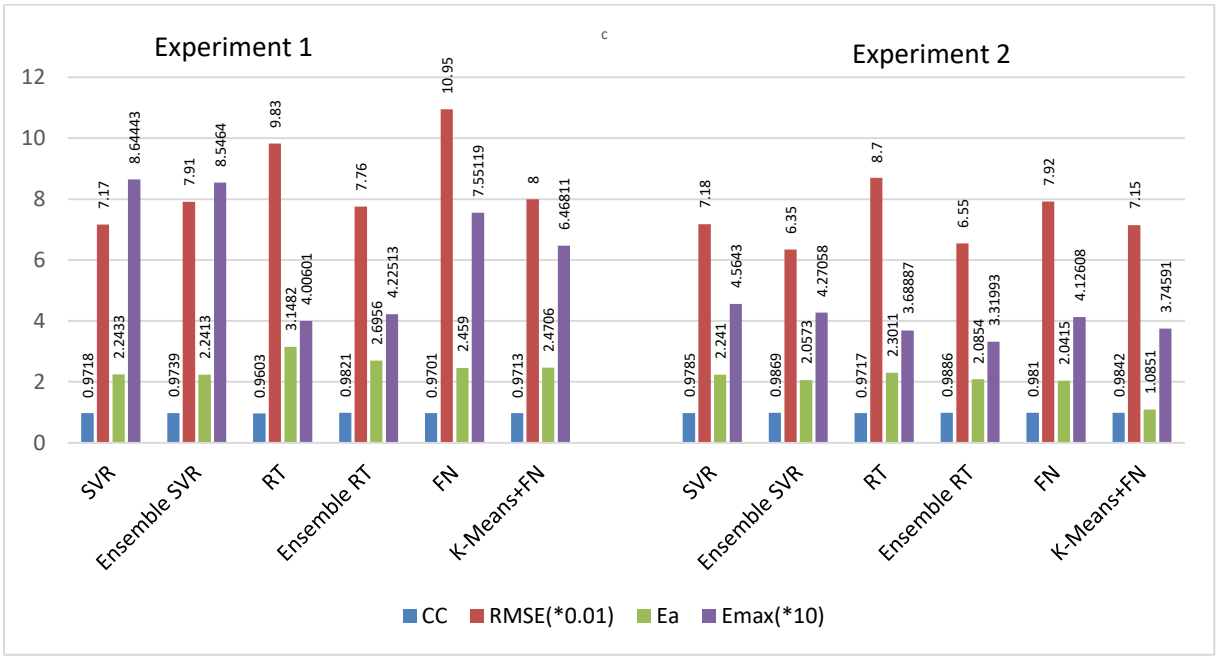


Figure 5.14. Performances of ML techniques (Experiments 1 and 2) on data set C for B_{ob} Prediction

Performances of the ML techniques for both experiments 1 and 2 on data set C for B_{ob} prediction are shown in Figure 5.14. Ensemble RT from experiment 2 gives the best performance with the largest *CC* and smallest *RMSE* and E_{max}. Going by the four performance measures, there is a tie between K-Means+FN and ensemble SVR from experiment 2 on data set C. The former has greater *CC* and lower *RMSE* while the latter has lower E_a and E_{max}. The performances of the ML techniques from experiment 2 improve virtually across all the statistical measures. Some empirical correlations have better or very competitive performances with some ML models from experiment 1 while all the ML techniques from experiment 2, except RT, are better than all the empirical correlations in at least 3 out of the 4 statistical measures.

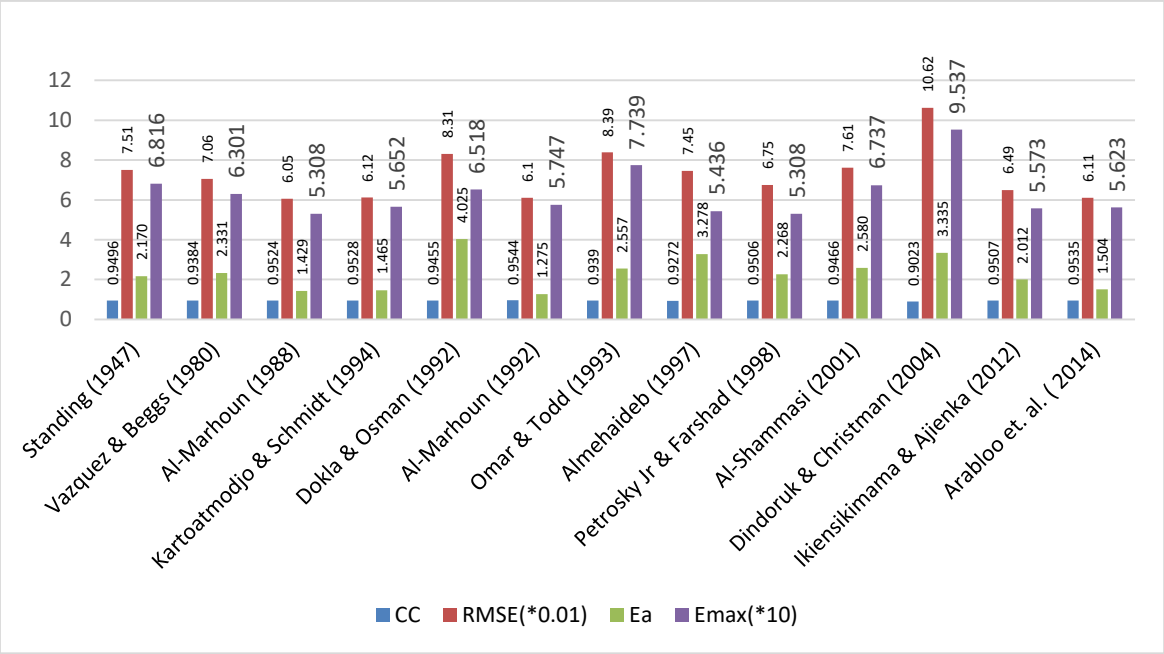


Figure 5.15. Performances of empirical correlations on data set D for B_{ob} prediction

The performances of the empirical correlations on data set D for B_{ob} prediction are shown in Figure 5.15. There are competitive leading performances among Al-Marhoun (1988), Kartoatmodjo & Schmidt (1994), Al-Marhoun (1992) and Arabloo et. al. (2014). The worst overall performance is given by Dindoruk & Christman (2004).

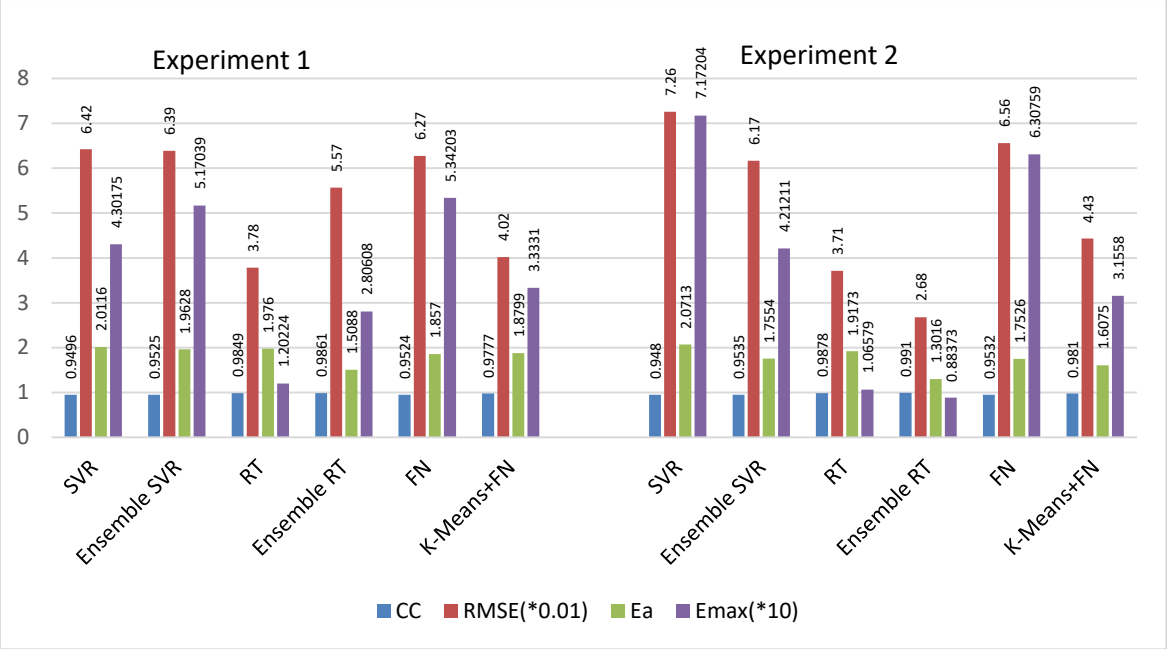


Figure 5.16. Performances of ML techniques (Experiments 1 and 2) on data set D for B_{ob} Prediction

The performances of ML techniques on data set D for both experiments 1 and 2 for B_{ob} prediction are shown in Figure 5.16. Ensemble RT from experiment 2 gives the best results with the highest *CC*, smallest *RMSE*, E_a and E_{max}. For some ML techniques in experiment 2, their respective methods from experiment 1 show competitive results with respect to some statistical measures. More details will be given on this in section 5.3.2. Some empirical correlations have competitive or better results than some ML techniques but all of them still have lower performances than a few ML techniques such as standalone RT, ensemble RT and K-Means+FN from both experiments 1 and 2.

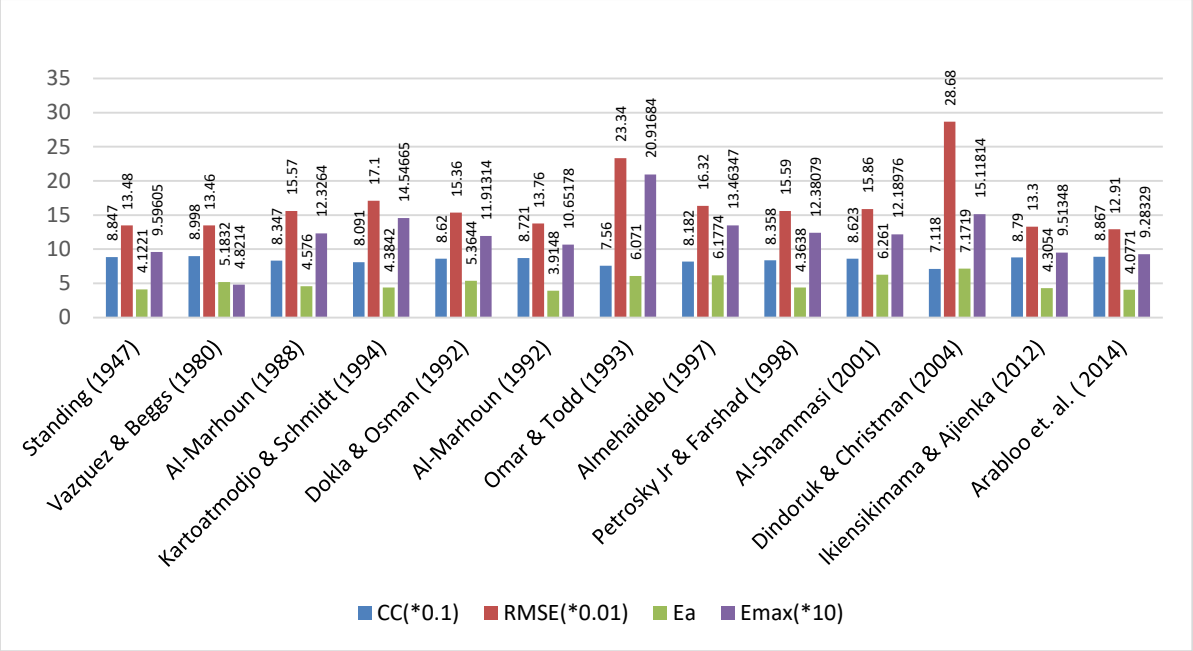


Figure 5.17. Performances of empirical correlations on data set E for B_{ob} prediction

The performances of empirical correlations for B_{ob} prediction using data set E are shown in Figure 5.17. Generally, many of these correlations perform poorly with the worst performance from Dindoruk & Christman (2004). The scale for *CC* has also been formatted to accommodate its low values. Clearly, these empirical correlations have not captured the uncertainties in this data set which is one of the problems associated with them vis-à-vis generalisation and consistency.

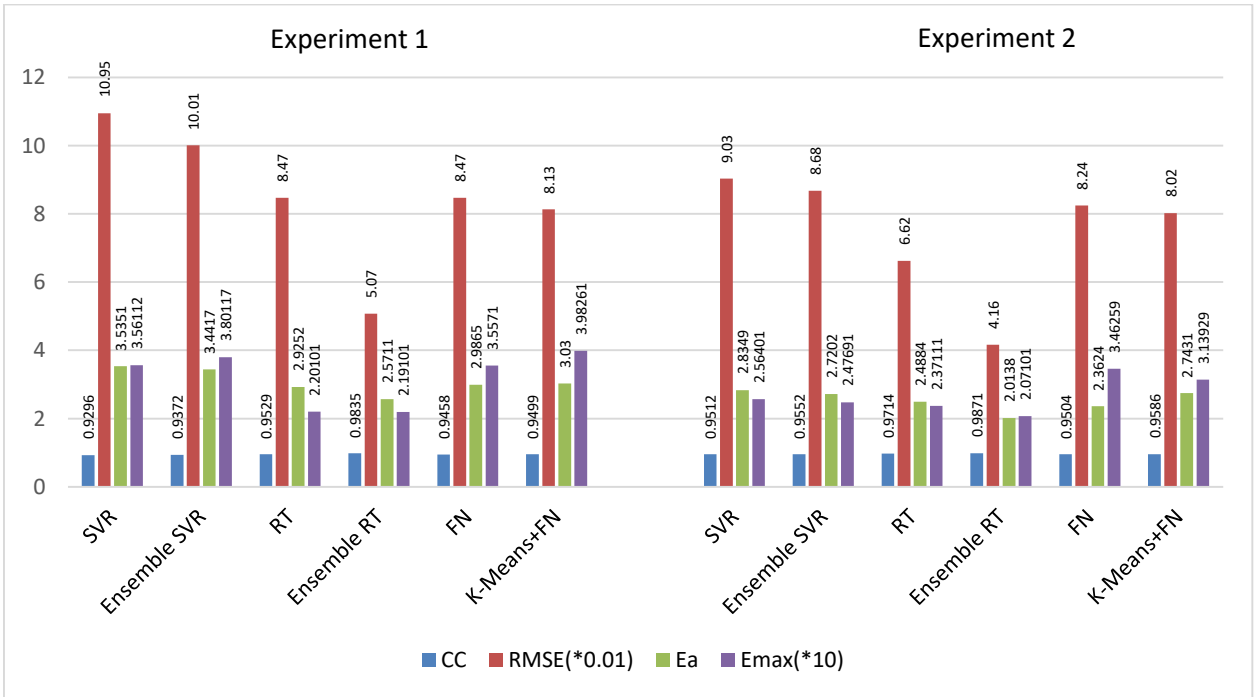


Figure 5.18. Performances of ML techniques (Experiments 1 and 2) on data set E for B_{ob} prediction

The performances of ML techniques from both experiment 1 and 2 on data set E for B_{ob} prediction are shown in Figure 5.18. Generally, the performances of the ML models show little deviation compared to the previous data sets. On the other hand, the empirical correlations show much clearer lower performance on this data set compared to data sets A, C and D. The overall best performance on this data set is given by ensemble RT leads with the largest *CC*, lowest *RMSE*, E_a and E_{max}. All ML models from both experiments 1 and 2 clearly perform better than all the empirical correlations in Table 5.17.

5.3.2. Trend and Comparative Analysis of Prediction of Oil Formation Volume Factor at Bubblepoint

Similar to the P_b prediction analysis, Tables 5.4 and 5.5 give a summary of the minimum and maximum values of the statistical measures for B_{ob} predictions from ML models of both experiments 1 and 2 and the comparing empirical correlations. The summary reports reflect performances on data sets C, D and E based on the given results under section 5.3.1 since ML models from experiments 1 and 2 are exclusive with data sets A and B respectively.

Table 5.4. Summary of Statistical Measures for B_{ob} Empirical Correlations

Correlation Method	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
Standing (1947)	0.8847 / 0.9552	0.075 / 0.135	2.170 / 4.122	68.158 / 95.961
Vazquez & Beggs (1980)	0.8998 / 0.9707	0.071 / 0.135	2.331 / 5.183	45.512 / 63.006
Al-Marhoun (1988)	0.8347 / 0.9629	0.061 / 0.156	1.429 / 4.576	53.075 / 123.264
Kartoatmodjo & Schmidt (1994)	0.8091 / 0.9675	0.061 / 0.171	1.465 / 4.384	56.518 / 145.467
Dokla & Osman (1992)	0.8620 / 0.9515	0.083 / 0.154	4.025 / 5.364	65.184 / 119.131
Al-Marhoun (1992)	0.8721 / 0.9674	0.061 / 0.138	1.275 / 3.915	57.467 / 106.518
Omar & Todd (1993)	0.7560 / 0.9647	0.084 / 0.233	2.557 / 6.071	77.392 / 209.168
Almehaideb (1997)	0.8182 / 0.9567	0.075 / 0.163	3.278 / 6.177	35.875 / 134.635
Petrosky Jr & Farshad (1998)	0.8358 / 0.9579	0.068 / 0.156	2.268 / 4.364	53.079 / 123.808
Al-Shammasi (2001)	0.8623 / 0.9714	0.071 / 0.159	2.580 / 6.261	56.638 / 121.898
Dindoruk & Christman (2004)	0.7118 / 0.9140	0.106 / 0.287	3.335 / 7.172	95.375 / 151.181
Ikiensikimama & Ajienka (2012)	0.8790 / 0.9555	0.065 / 0.133	2.012 / 4.305	55.728 / 95.135
Arabloo et. al. (2014)	0.8867 / 0.9666	0.061 / 0.129	1.504 / 4.077	51.716 / 92.833

Table 5.5. Summary of Statistical Measures for Testing B_{ob} Prediction with ML techniques

Experiment 1	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
SVR	0.9296 / 0.9718	0.064 / 0.110	2.012 / 3.535	35.611 / 86.444
Ensemble SVR	0.9372 / 0.9739	0.064 / 0.100	1.963 / 3.442	38.012 / 85.464
RT	0.9529 / 0.9849	0.038 / 0.098	1.976 / 3.148	12.022 / 40.060
Ensemble RT	0.9821 / 0.9861	0.051 / 0.078	1.509 / 2.696	21.910 / 42.251
FN	0.9458 / 0.9701	0.063 / 0.110	1.857 / 2.987	35.571 / 75.512
K-Means+FN	0.9499 / 0.9777	0.040 / 0.081	1.880 / 3.030	33.331 / 64.681
Experiment 2				
SVR	0.9480 / 0.9785	0.072 / 0.090	2.071 / 2.835	25.640 / 71.720
Ensemble SVR	0.9535 / 0.9869	0.062 / 0.087	1.755 / 2.720	24.769 / 42.706
RT	0.9714 / 0.9878	0.037 / 0.087	1.917 / 2.488	10.658 / 36.889
Ensemble RT	0.9871 / 0.9910	0.027 / 0.066	1.302 / 2.085	8.837 / 33.199
FN	0.9504 / 0.9810	0.066 / 0.082	1.753 / 2.362	34.626 / 63.076
K-Means+FN	0.9586 / 0.9842	0.044 / 0.080	1.085 / 2.743	31.393 / 37.459

From the summary reports of Tables 5.4 and 5.5, the following points can be inferred about the performances of the ML models and empirical correlations.

- All ML models from both experiments 1 and 2 have higher minimum and maximum CC than all the empirical correlations except FN from experiment 1 which slightly trails Al-Shammasi (2001) and Vazquez & Beggs (1980) for maximum CC. Higher CC values of the ML models imply better matching of their predicted values with the targets. Lower minimum CC of the empirical correlations implies their higher likelihood mismatch prediction.
- While some empirical correlations have their minimum RMSE competitive with respective results of some ML models, all the ML models from both experiments have lower maximum RMSE.
- Though few empirical correlations have smaller minimum E_a than some ML models from both experiments, all empirical correlations have higher maximum E_a than all the ML models.
- All ML models have smaller minimum E_{max} than all the comparing empirical correlations except the result of ensemble SVR from experiment which is higher than the corresponding measure for Almehaideb (1997). However, Almehaideb (1997) has higher maximum E_{max} than the ensemble SVR and all other ML models. All the comparing empirical correlations

have poorer maximum E_{max}, (i.e. higher values), than the ML models except for Vazquez & Beggs (1980) which has lower maximum E_{max} than some ML models.

Overall, some ML models (especially SVR, ensemble SVR, ensemble RT, FN and K-Means+FN from experiment 2) can be concluded to have better performances than all the empirical correlations. Whenever a correlation is competitive with these ML models with respect to a statistical measure, its other performance measures are lower.

Table 5.6 also compares the performances of only the ML techniques for B_{ob} prediction. Similar to P_b prediction, there is a significant improvement on the performances from the models developed with data set B over the models from data set A. Unlike the P_b prediction, the overall best performance can be unequivocally be attributed to ensemble RT from experiment 2.

Likewise, similar trends to P_b prediction have been shown by the evaluated empirical correlations B_{ob} prediction. The empirical correlations show large deviations in the results for data set E compared to other data sets (A-D). It is observed that data set E has the maximum γ_{API} of 115 which is the largest among all the data sets. On the other hand, most empirical correlations have smaller deviations between data sets A and B, or even better performance on the latter, compared to what can be observed for empirical correlations on data sets A and B under P_b prediction.

Comparing the ranges of GOR of the data sets with the performances of different models, the empirical correlations perform better with data set C, where the maximum GOR is 3617.27 SCF/STB and maximum γ_{API} is 63.7, than data set E where the maximum GOR is maximum 2500 scf/STB and maximum γ_{API} of 115. Hence, the GOR might have not influenced the performances of the empirical models for the of B_{ob} as much as γ_{API}. In contrary, the performances of the ML models from experiment 2 are quite comparative on the two data sets.

Table 5.6. Deductive Comparison of ML Technique for Oil FVF Prediction

	Data Sets				
Methods	A	B	C	D	E
SVR vs Ensemble SVR	Ensemble SVR	Ensemble SVR	Ensemble SVR	Ensemble SVR	Ensemble SVR
RT vs Ensemble RT	Ensemble RT	Ensemble RT	Ensemble RT	Ensemble RT	Ensemble RT
FN vs K-Means+FN	K-Means+FN	K-Means+FN	K-Means+FN	K-Means+FN	K-Means+FN

To corroborate the inferences and the summary reports of Tables 5.4 and 5.5, average of the statistical measures for the B_{ob} prediction from the selected empirical correlations and the six ML models are shown in Figure 5.19 and 5.20.

All the ML models from experiments 1 and 2 have higher average *CC* than all the empirical correlations, indicating averagely, better matching between their predicted outputs and the targets. With respect to *RMSE*, E_a and E_{max}, all the ML models from experiment 2 averagely outperform all the empirical correlations. Also, all the ML models from experiment 1 have lower average E_a than all the empirical correlations while only very few of the correlations have competitive average $RMSE$ and E_{max} with the ML models from experiment 1.

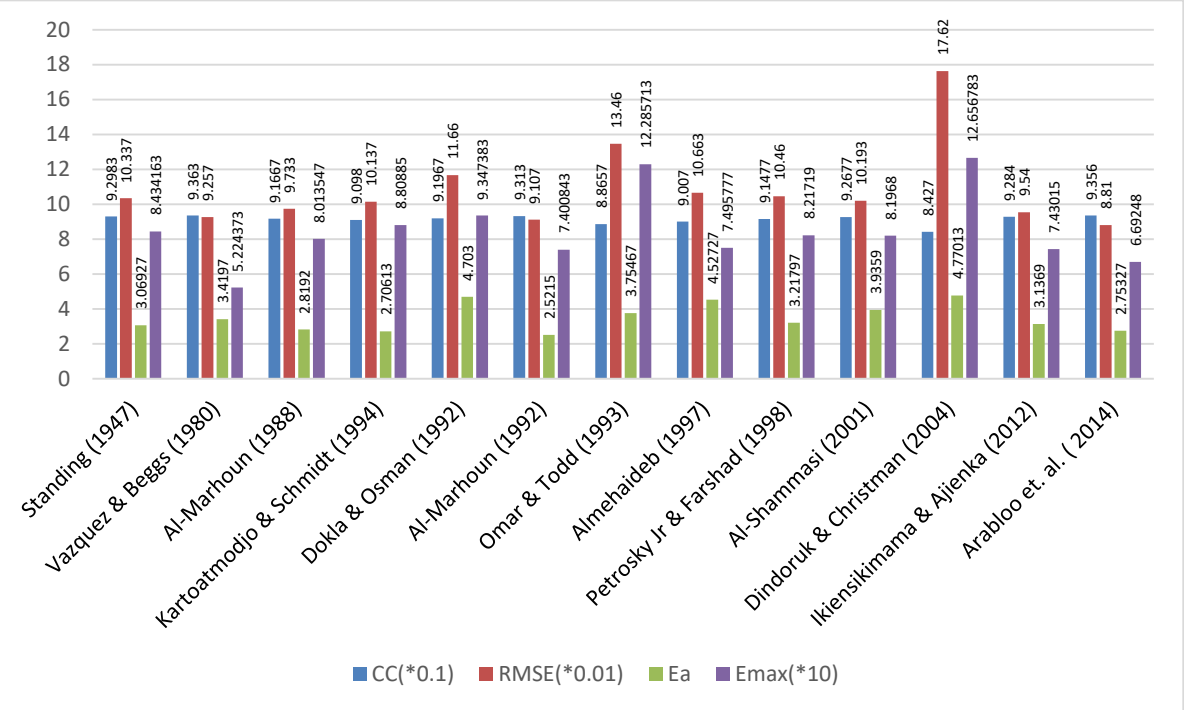


Figure 5.19. Average of Statistical Measures for B_{ob} Empirical Correlations

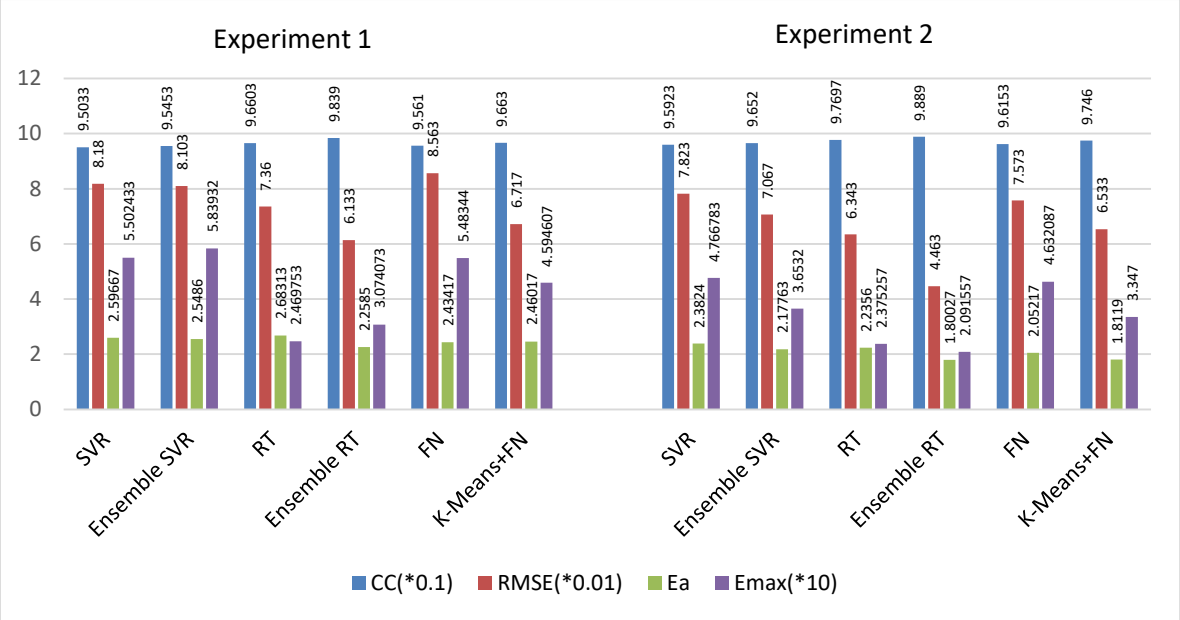


Figure 5.20. Average of Statistical Measures for B_{ob} Prediction with ML techniques

5.4. Prediction of Dead Oil Viscosity

One of the most difficult PVT properties to predict is μ_{od}. Different correlating variables have been used in additional to the two common correlating properties (γ_{API}, T). Four of these functional forms which do not use hydrocarbon components have been considered.

$$\mu_{od} = f(\gamma_{API}, T)$$

$$\mu_{od} = f(\gamma_{API}, T, R_s)$$

$$\mu_{od} = f(\gamma_{API}, T, P_b)$$

$$\mu_{od} = f(\gamma_{API}, T, R_s, P_b)$$

5.4.1. Results and Error Analysis for Dead Oil Viscosity

Based on the four functional forms of μ_{od}, different models for the six considered ML techniques have been developed and then subsequently tested on different data sets. Figures 5.21 to 5.25 show the results of these experimental setups.

The evaluating correlations have been categorised according to their functional forms in Table 5.7

Table 5.7. Categorisation of the Evaluated μ_{od} Correlations Based on their Functional Forms

Correlation Methods	Functional Form
Beggs & Robinson (1975); Glasø (1980); Naseri et al. (2005); Kartoatmodjo & Schmidt (1994); Petrosky Jr and Farshad (1998); Labedi (1992); Elsharkawy and Alikhan (1999)	γ_{API}, T
Dindoruk and Christman (2004)	$\gamma_{API}, T, R_{sb}, P_b$

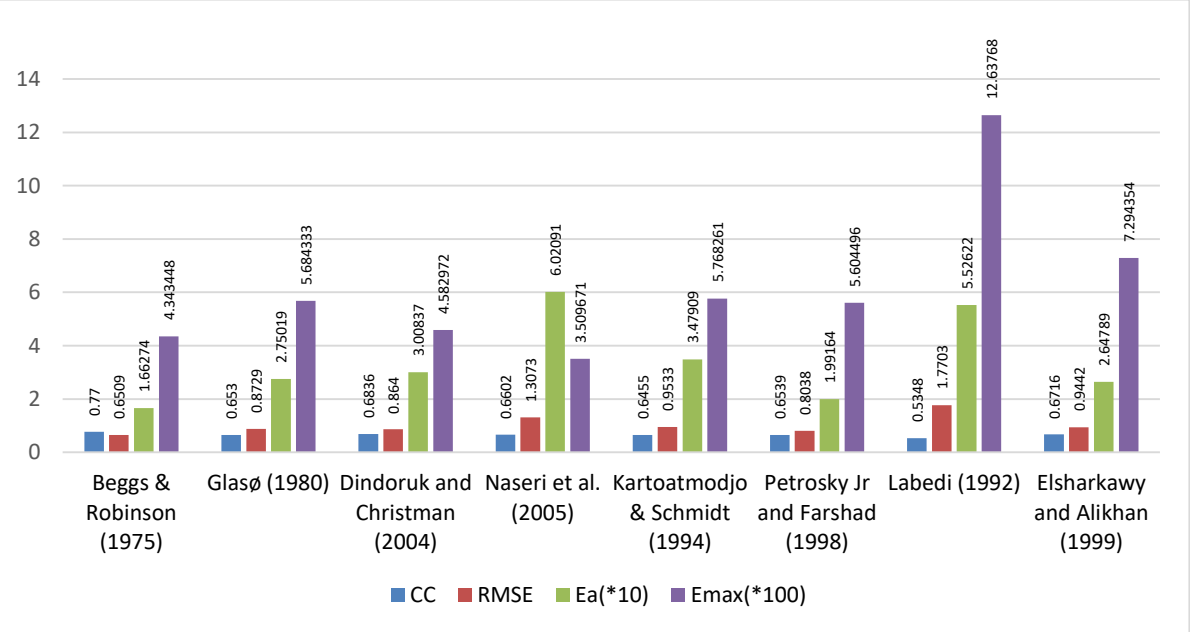


Figure 5.21. Performance of μ_{od} Empirical Correlations on data set F

The performances of the μ_{od} empirical correlations on data set F are shown in Figure 5.21. Beggs & Robinson (1975) shows the best performance among these correlations with the lowest *RMSE* and E_a. High E_{max} across all the evaluated empirical correlations show that they have largely over-predicted one or more data points. Labedi (1992) shows the worst performance for the μ_{od} prediction among these methods on this data set.

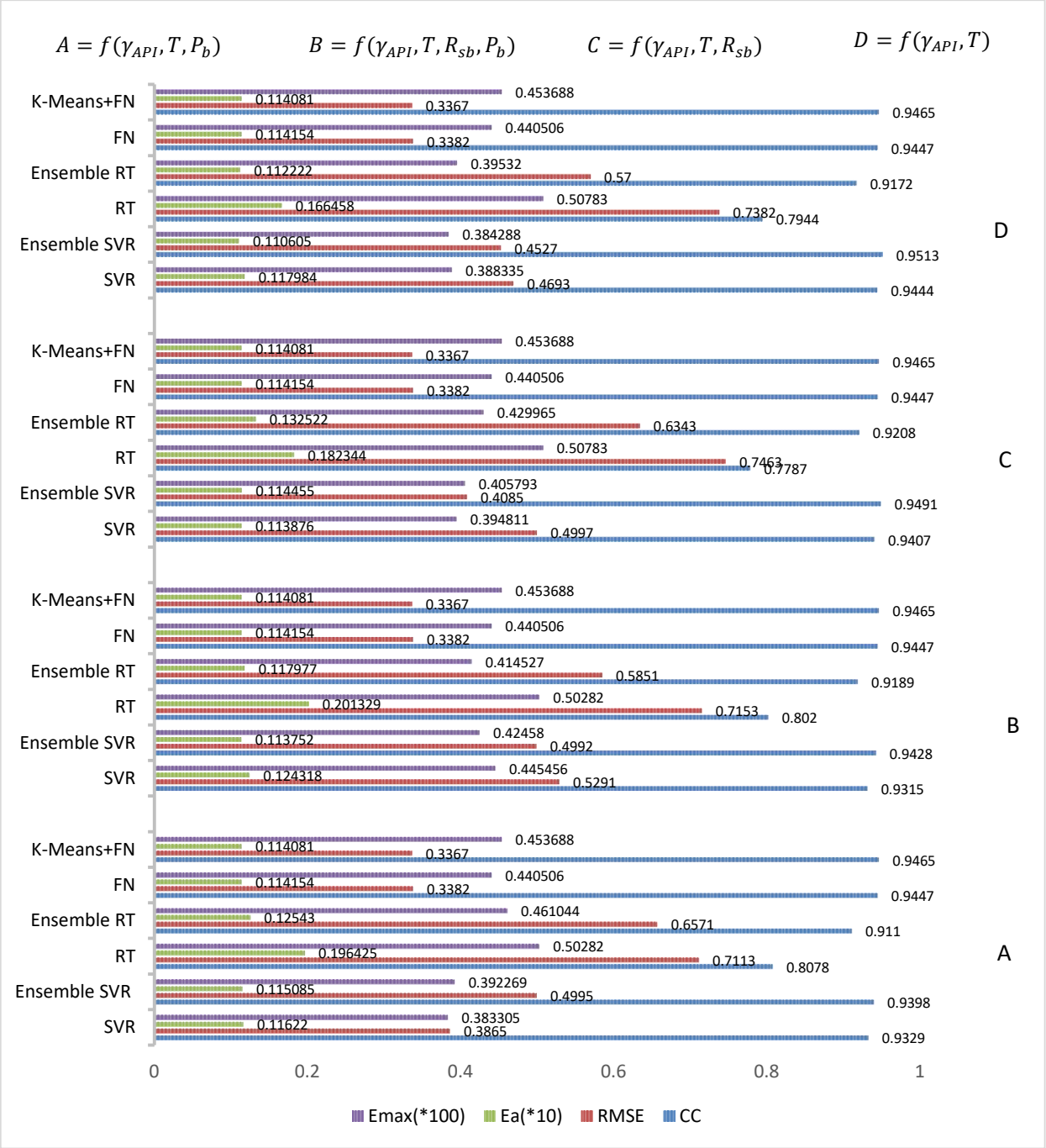


Figure 5.22. Performance of ML techniques for μ_{od} prediction with data set F

The performances of the ML techniques on data set F for μ_{od} prediction are shown in Figure 5.22. The standalone FN and hybrid K-Means+FN show interesting pattern and behaviour, giving same results across all the functional forms. Since the variables in the functional form $f(\gamma_{API}, T)$ are a subset of other functional forms, it means that the FN has eliminated all other variables, correlating μ_{od} with only γ_{API} and T. This is in fact in line with the fact that many empirical correlations have been developed based on these two variables. The hybrid K-Means+FN has the least *RMSE* across all functional forms for all methods. The overall best method lies among the ensemble SVR for the

functional form $f(\gamma_{API}, T, P_b)$ and K-Means+FN. The ensemble SVR with the functional form $f(\gamma_{API}, T)$ has the highest CC.

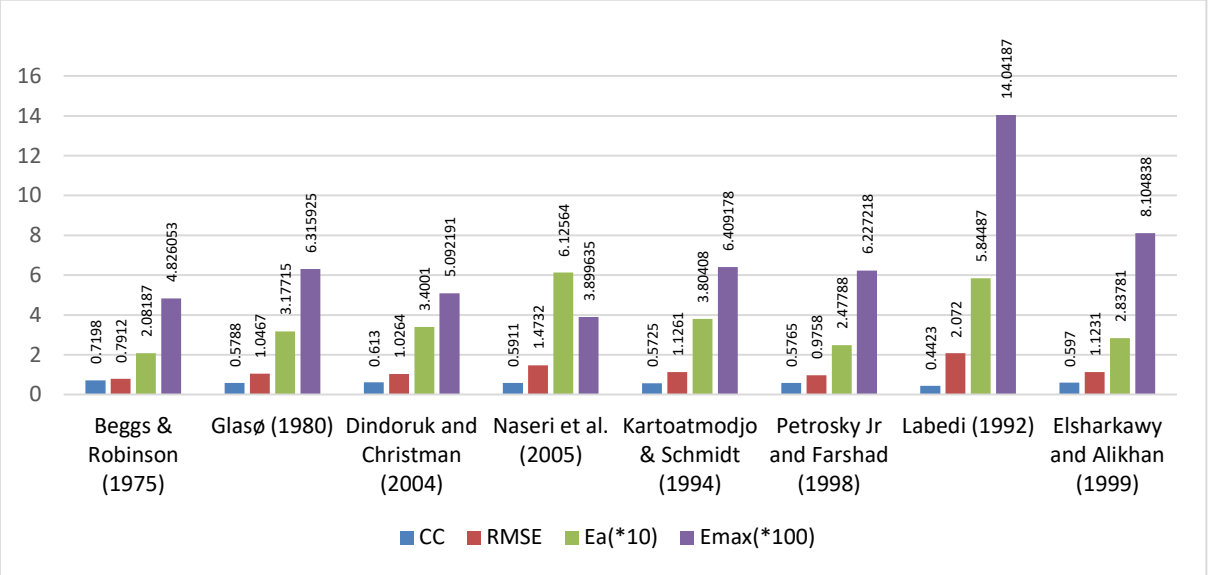


Figure 5.23. Performance of μ_{od} Empirical Correlations on data set G

Performances of some selected empirical correlations for μ_{od} prediction on data set G are shown in Figure 5.23. Among these empirical correlations, Beggs & Robinson (1975) has the best results with the highest CC, lowest RMSE and E_a, though its E_{max} is the second highest which indicates almost four times over or under prediction of one or more data points. The worst performance is demonstrated by Labedi (1992) with unacceptably high E_{max}.

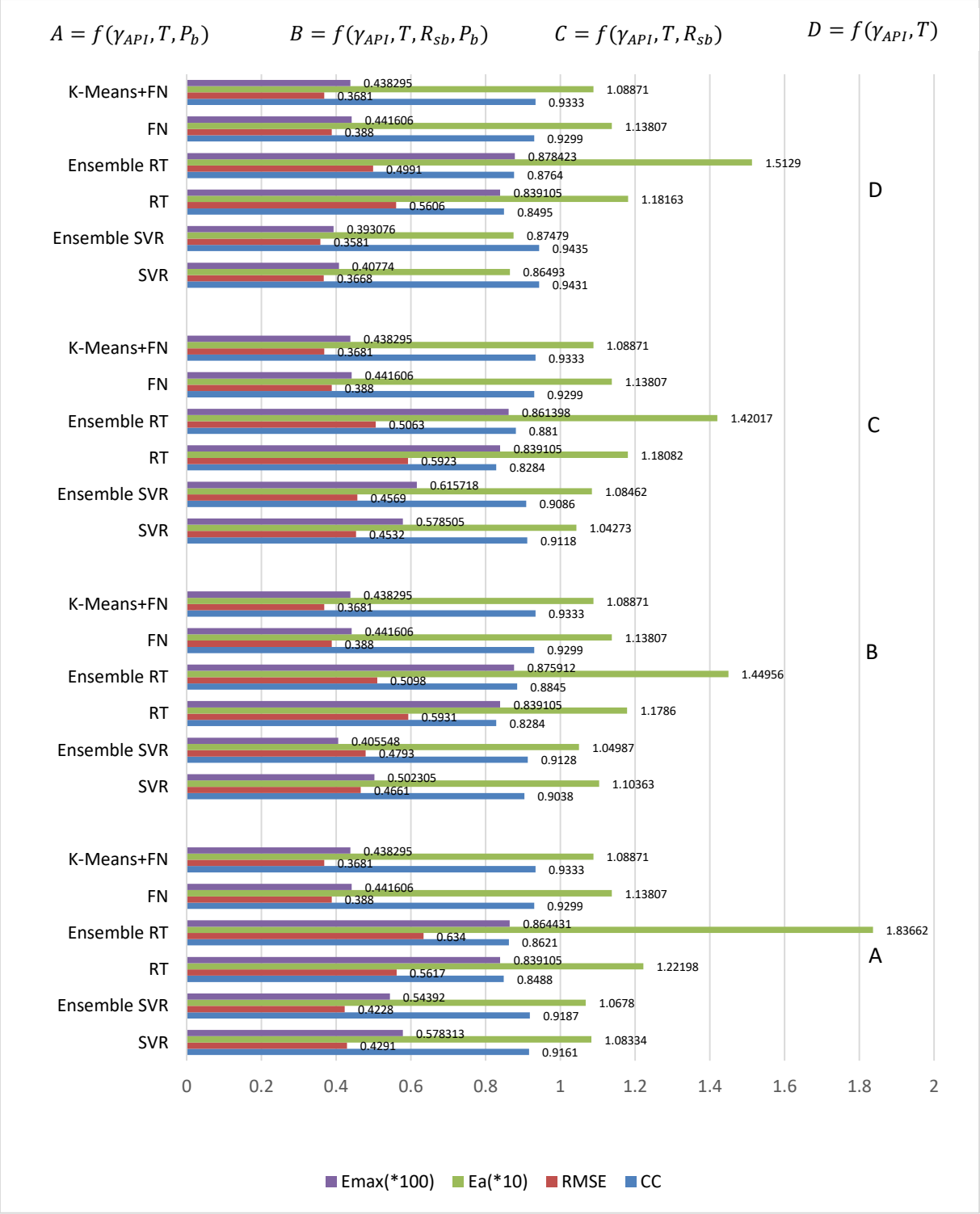


Figure 5.24. Performance of ML techniques for μ_{od} prediction with data set G

The performances of ML techniques for μ_{od} prediction using data set G are presented in Figure 5.24. The overall best performance is given by ensemble SVR for functional form $f(\gamma_{API}, T)$ with the highest CC, minimum *RMSE* and E_{max}. The standalone SVR for the same functional form also

shows very good results. Similar to data set F, the results for the standalone FN and hybrid K-Means+FN are the same across all the functional forms. Both standalone RT and ensemble RT have not shown impressive performances for μ_{od} prediction with both data sets F and G.

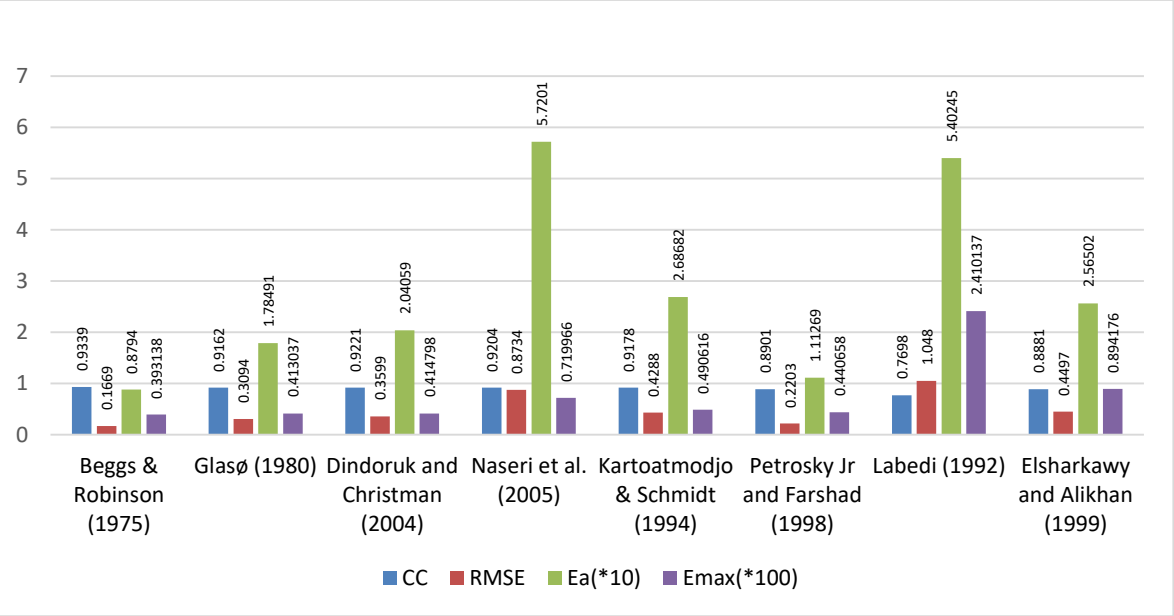


Figure. 5.25. Performance of μ_{od} Empirical Correlations on data set H

The results of the selected empirical correlations for μ_{od} prediction with data set H are shown in Figure 5.25. Beggs & Robinson (1975) with the functional form of $f(\gamma_{API}, T)$ has the best performance with the largest *CC* and smallest *RMSE*, E_a and E_{max}. It is noted that additional correlating variable introduced by Dindoruk and Christman (2004) in their correlation has not improved its results significantly with the performance trailing the correlations of Beggs & Robinson (1975) and Glasø (1980) which give first and second best performances respectively in Figure 5.25.

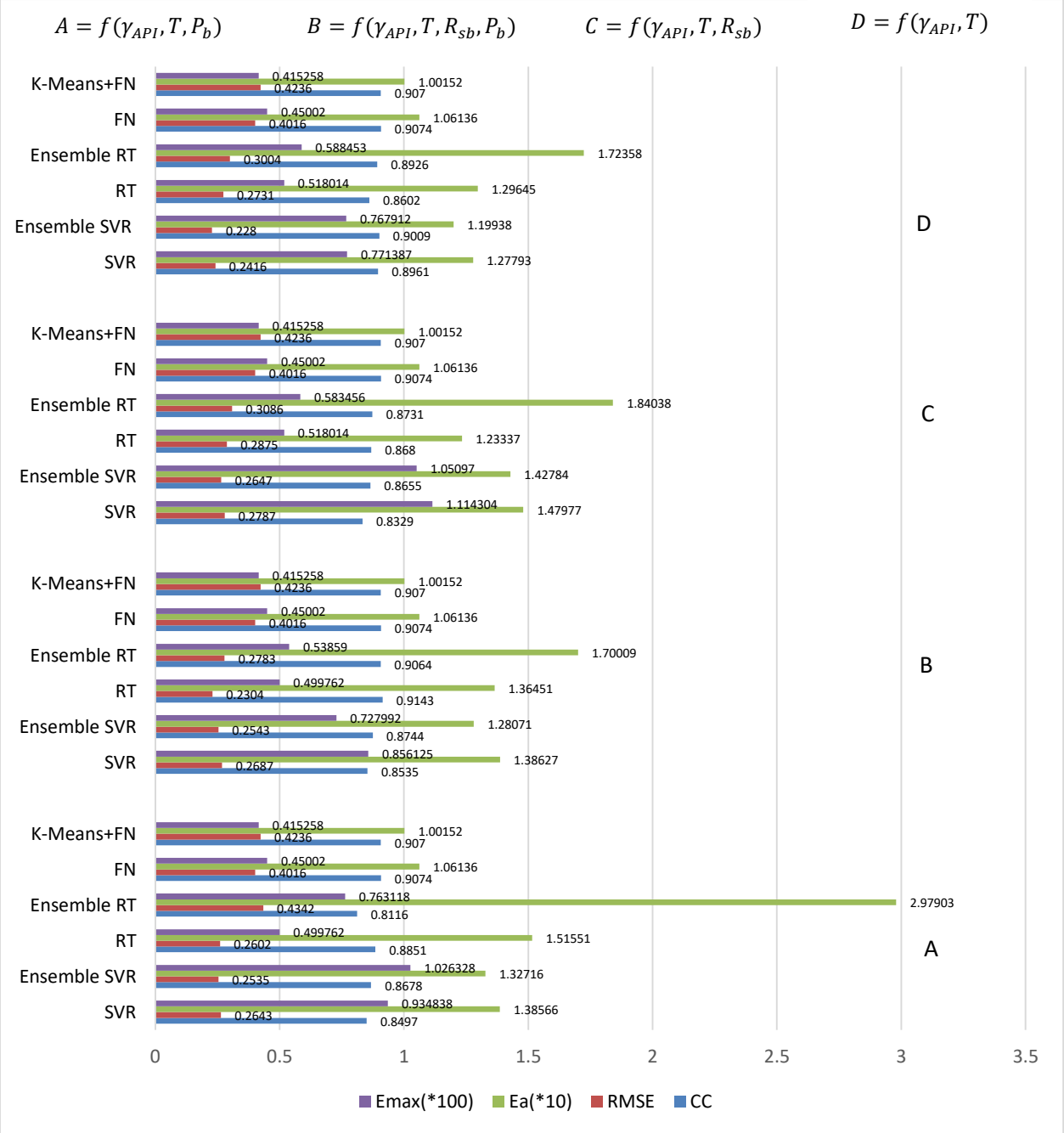


Figure 5.26. Performance of ML techniques for μ_{od} prediction with data set H

The results for μ_{od} prediction using ML models on data set H is presented in Figure 5.26. Generally, the performances of the ML techniques for μ_{od} prediction with all the functional forms are not as good as their predictions for the data sets F and G. FN and K-Means+FN show most reliable and consistent results among the ML methods where FN gives the largest CC while K-Means+FN gives the smallest E_a and E_{max}

Comparing the results of the empirical correlations with ML techniques for this data set, Beggs & Robinson (1975) outperforms all the ML techniques though its performances on the previous two

data sets were comparably lower than many other ML techniques. Performances of some empirical correlations are very competitive with many ML techniques across all functional forms on this data set.

5.4.2. Trend and Comparative Analysis of Dead Oil Viscosity Prediction

The performances of the ML techniques for the prediction of dead oil viscosity are summarized in Table 5.8. It can be noticed that most leading techniques in each category perform well with the functional form of $f(\gamma_{API}, T)$.

For data set H, it is observed that the best correlation in this testing case, Beggs & Robinson (1975), outperforms all ML techniques including the FN and K-Means+FN which have leading performances among the ML techniques on this data set. It can be observed that the statistics of data sets F and G overlap more that the statistics for data set H. Hence, the possible reason while the performances of the models on the two data sets are similar.

In the same vein, the additional correlating parameters have not convincingly improved the performances of most ML techniques for the prediction of μ_{od}. The ensemble SVR with the functional form $f(\gamma_{API}, T)$ has best performance among the ML techniques and all empirical correlations for both data sets F and G. In the same vein, the leading method among the empirical correlations, Beggs & Robinson (1975), uses the same functional form. Consequently, additional correlating parameters might be unnecessary. In fact, FN and K-Means+FN have intelligently suppressed all other input variables (R_{sb} and P_b), utilizing only γ_{API} and T.

Table 5.8. Deductive Comparison of ML Technique for Dead Oil Viscosity Prediction

	Data Sets		
Methods	F	G	H
SVR vs Ensemble SVR	Ensemble SVR [$f(\gamma_{API}, T)$]	Ensemble SVR [$f(\gamma_{API}, T)$]	SVR, $f(\gamma_{API}, T, R_{sb})$
RT vs Ensemble RT	Ensemble RT [$f(\gamma_{API}, T)$]	Tie, $f(\gamma_{API}, T, R_{sb}, P_b)$	RT, $f(\gamma_{API}, T, R_{sb}, P_b)$
FN vs K-Means+FN	K-Means+FN [$f(\gamma_{API}, T)$]	K-Means+FN, $f(\gamma_{API}, T)$	Tie, $f(\gamma_{API}, T)$

Tables 5.9 and 5.10 give the minimum and maximum values of the statistical measures of the empirical correlations for μ_{od} prediction. As previously discussed, Both FN and K-Means+FN have same results for all the functional forms. This implies that only γ_{API} and T which are subset of other functional forms have been found to be adequate and true representation of the target, μ_{od}. Hence, the suppression of other input variables.

On the other hand, other four ML models appear to correlate the target with other input to an extent. For simplicity, the results of the ML models in Table 5.5.11 with functional form $f(\gamma_{API}, T)$ will be compared with the empirical correlations in Table 5.5.10 and the following inferences can be made.

- With respect to the minimum and maximum CC, Four ML techniques (SVR, ensemble SVR, FN and K-Means+FN)) outperform all the empirical correlations with much higher values, indicating better matching between their predictions and the targets.
- While the minimum values of RMSE, E_a and E_{max} for some empirical correlations are competitive with the corresponding values for the aforementioned four ML models, their maximum statistical measure values are much higher, indicating poorer predictions.

In the same vein, Table 5.5.12 and 5.5.13 corroborate the points above with the four mentioned ML techniques having much higher average *CC*, lower average *RMSE*, E_a and E_{max}. In fact the average performances of other ML techniques (RT and ensemble RT) with the same functional form are also better than all the empirical correlations.

Table 5.9. Summary of Statistical Measures for μ_{od} Empirical Correlations

	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
Beggs & Robinson (1975)	0.7198 / 0.9339	0.167 / 0.791	8.794 / 20.819	39.314 / 482.605
Glasø (1980)	0.5788 / 0.9162	0.309 / 1.047	17.849 / 31.771	41.304 / 631.593
Dindoruk and Christman (2004)	0.6130 / 0.9221	0.360 / 1.026	20.406 / 34.001	41.480 / 509.219
Naseri et al. (2005)	0.5911 / 0.9204	0.873 / 1.473	57.201 / 61.256	71.997 / 389.964
Kartoatmodjo & Schmidt (1994)	0.5725 / 0.9178	0.429 / 1.126	26.868 / 38.041	49.062 / 640.918
Petrosky Jr and Farshad (1998)	0.5765 / 0.8901	0.220 / 0.976	11.127 / 24.779	44.066 / 622.722
Labedi (1992)	0.4423 / 0.7698	1.048 / 2.072	54.025 / 58.449	241.014 / 1404.18
Elsharkawy and Alikhan (1999)	0.5970 / 0.8881	0.450 / 1.123	25.650 / 28.378	89.418 / 810.484

Table 5.10. Summary of Statistical Measures for Testing u_{od} Prediction with ML techniques

$f(\gamma_{API}, T, P_b)$	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
SVR	0.8497 / 0.9329	0.264 / 0.429	10.833 / 13.857	38.330 / 93.484
Ensemble SVR	0.8678 / 0.9398	0.253 / 0.500	10.678 / 13.272	39.227 / 102.633
RT	0.8078 / 0.8851	0.260 / 0.711	12.220 / 19.642	49.976 / 83.911
Ensemble RT	0.8116 / 0.9110	0.434 / 0.657	12.543 / 29.790	46.104 / 86.443
FN	0.9074 / 0.9447	0.338 / 0.402	10.614 / 11.415	44.051 / 45.002
K-Means+FN	0.9070 / 0.9465	0.337 / 0.424	10.015 / 11.408	41.526 / 45.369
$f(\gamma_{API}, T, R_{sb}, P_b)$				
SVR	0.8535 / 0.9315	0.269 / 0.529	11.036 / 13.863	44.546 / 85.613
Ensemble SVR	0.8744 / 0.9428	0.254 / 0.499	10.499 / 12.807	40.555 / 72.799
RT	0.8020 / 0.9143	0.230 / 0.715	11.786 / 20.133	49.976 / 83.911
Ensemble RT	0.8845 / 0.9189	0.278 / 0.585	11.798 / 17.001	41.453 / 87.591
FN	0.9074 / 0.9447	0.338 / 0.402	10.614 / 11.415	44.051 / 45.002
K-Means+FN	0.9070 / 0.9465	0.337 / 0.424	10.015 / 11.408	41.526 / 45.369
$f(\gamma_{API}, T, R_{sb})$				
SVR	0.8329 / 0.9407	0.279 / 0.500	10.427 / 14.798	39.481 / 111.430
Ensemble SVR	0.8655 / 0.9491	0.265 / 0.457	10.846 / 14.278	40.579 / 105.097
RT	0.7787 / 0.8680	0.288 / 0.746	11.808 / 18.234	50.783 / 83.911
Ensemble RT	0.8731 / 0.9208	0.309 / 0.634	13.252 / 18.404	42.996 / 86.140
FN	0.9074 / 0.9447	0.338 / 0.402	10.614 / 11.415	44.051 / 45.002
K-Means+FN	0.9070 / 0.9465	0.337 / 0.424	10.015 / 11.408	41.526 / 45.369
$f(\gamma_{API}, T)$				
SVR	0.8961 / 0.9444	0.242 / 0.469	8.649 / 12.779	38.834 / 77.139
Ensemble SVR	0.9009 / 0.9513	0.228 / 0.453	8.748 / 11.994	38.429 / 76.791
RT	0.7944 / 0.8602	0.273 / 0.738	11.816 / 16.646	50.783 / 83.911
Ensemble RT	0.8764 / 0.9172	0.300 / 0.570	11.222 / 17.236	39.532 / 87.842
FN	0.9074 / 0.9447	0.338 / 0.402	10.614 / 11.415	44.051 / 45.002
K-Means+FN	0.9070 / 0.9465	0.337 / 0.424	10.015 / 11.408	41.526 / 45.369

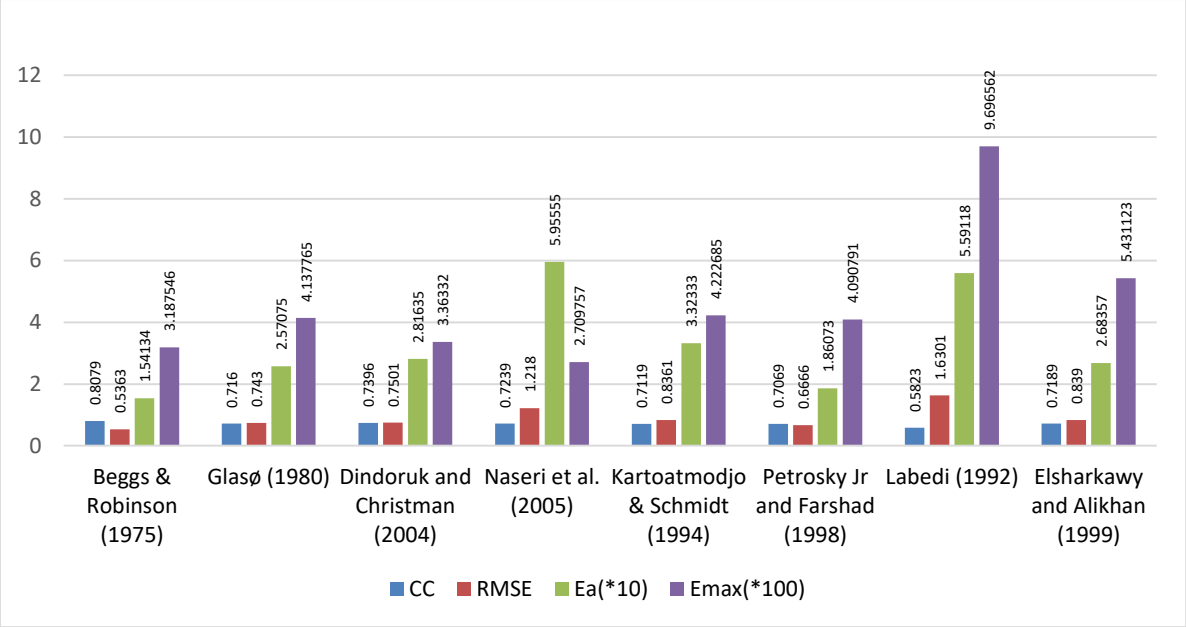


Figure 5.27. Average of Statistical Measures for μ_{od} Empirical Correlations

5.5. Prediction of Saturated Oil Viscosity

Four different functional forms have been examined for μ_{ob} prediction. These are as follow.

$$\mu_{ob} = f(\gamma_g, R_s, \gamma_o, T)$$
$$\mu_{ob} = f(\gamma_g, R_s, \gamma_{API}, T)$$
$$\mu_{ob} = f(\gamma_{API}, \mu_{od}, P_b)$$
$$\mu_{ob} = f(\mu_{od}, R_s)$$

5.5.1. Results and Error Analysis for Saturated Oil Viscosity

Different models have been developed for all the six ML techniques for the prediction of μ_{ob} based on the four functional forms under consideration. The results of the ML techniques are compared with some empirical correlations which use these functional forms. The ML models and the selected empirical correlations have been tested on different data sets to examine their generalisation and accuracy capabilities.

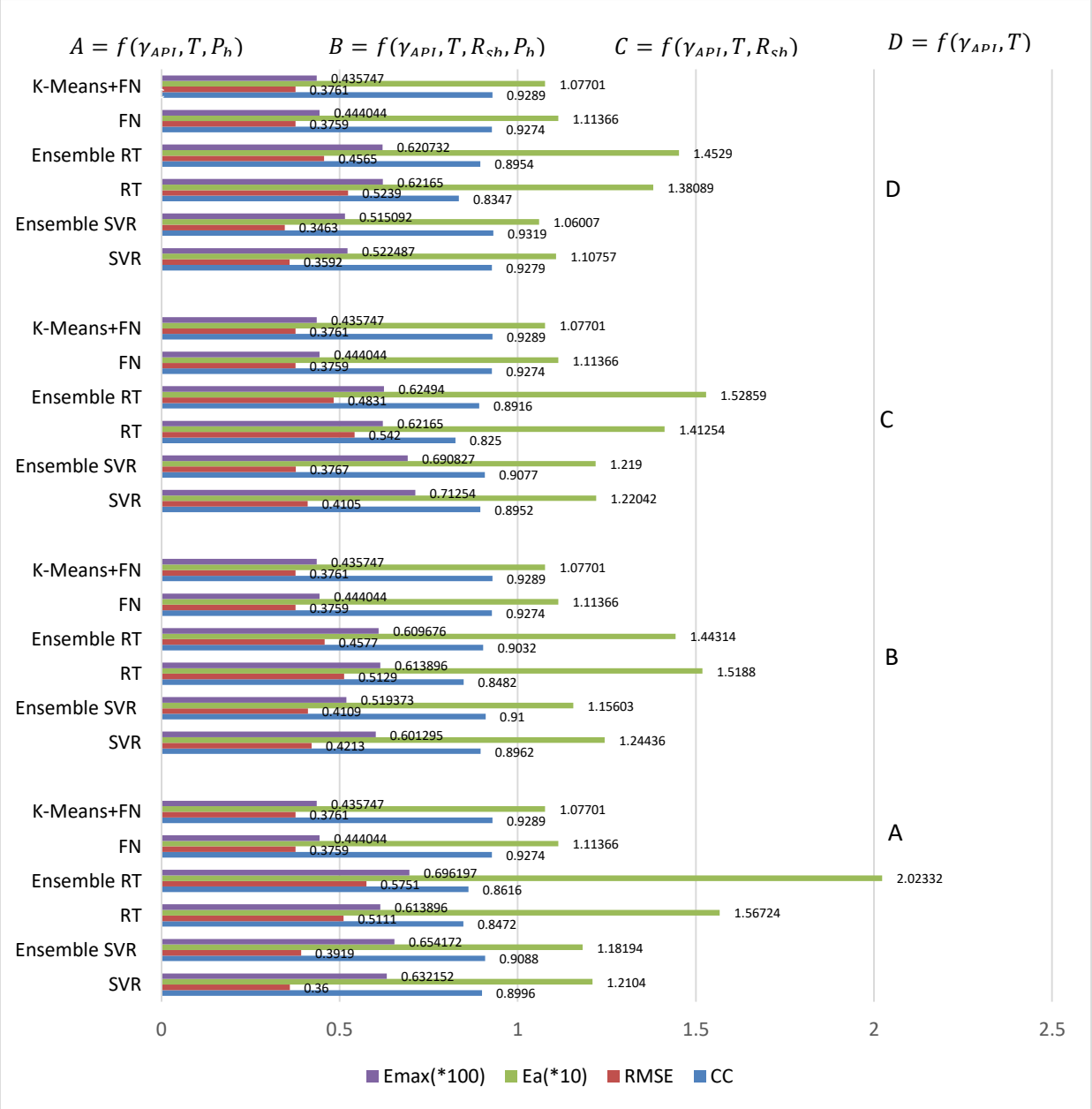


Figure 5.28. Average of Statistical Measures for Testing u_{od} Prediction with ML techniques

Table 5.11. Categorisation of the Evaluated μ_{ob} Correlations Based on their Functional Forms

Correlation Methods	Functional Form
Chew & Connally (1959); Vazquez & Beggs (1980); Elsharkawy & Alikhan (1999); Al-Khafaji et. al. (1987); Dindoruk & Christman(2004)	μ_{od}, R_{sb}
Khan et. al. (1987)	$\gamma_g, R_{sb}, \gamma_o, T$
Almehaideb (1997)	$\gamma_g, R_{sb}, \gamma_{API}, T$
Labedi (1992)	$\gamma_{API}, \mu_{od}, P_b$

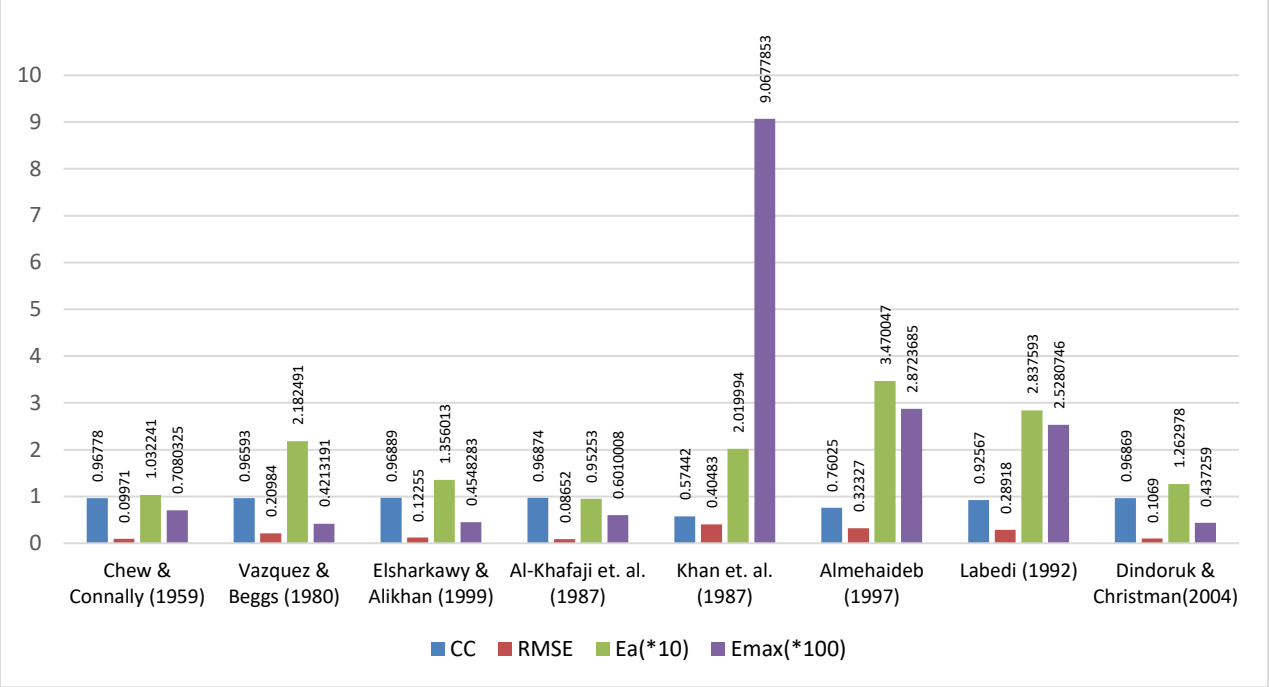


Figure 5.29. Performance of μ_{ob} Empirical Correlations on data set F

The performance of empirical correlations for the μ_{ob} prediction on data set F is shown in Figure 5.29. Among the empirical correlations, Al-Khafaji et. al. (1987) has the best performance with competitive value of *CC* and lowest error values for *RMSE* and E_a though its E_{max} is higher than the minimum value from Vazquez & Beggs (1980). However, other evaluating metrics do not favour Vazquez & Beggs (1980). The worst performance is given by Khan et. al. (1987) with a very large value of E_{max} which indicates huge over-prediction for some data point(s) and highest *RMSE*.

The performances of all the ML techniques for μ_{ob} prediction on data set F are shown in Figure 5.30. The overall best performance lies with the hybrid K-Means+FN for the functional form $f(\mu_{od}, R_{sb})$ with smallest *RMSE* and E_{max}, and highest CC. The standalone FN for this functional form shows the next best result which is competitive with the hybrid model. Good results are also shown by most ML techniques with the functional forms $f(\gamma_{API}, \mu_{od}, P_b)$ and $f(\mu_{od}, R_{sb})$. The RT of the functional form $f(\gamma_g, R_{sb}, \gamma_o, T)$ shows poor performance.

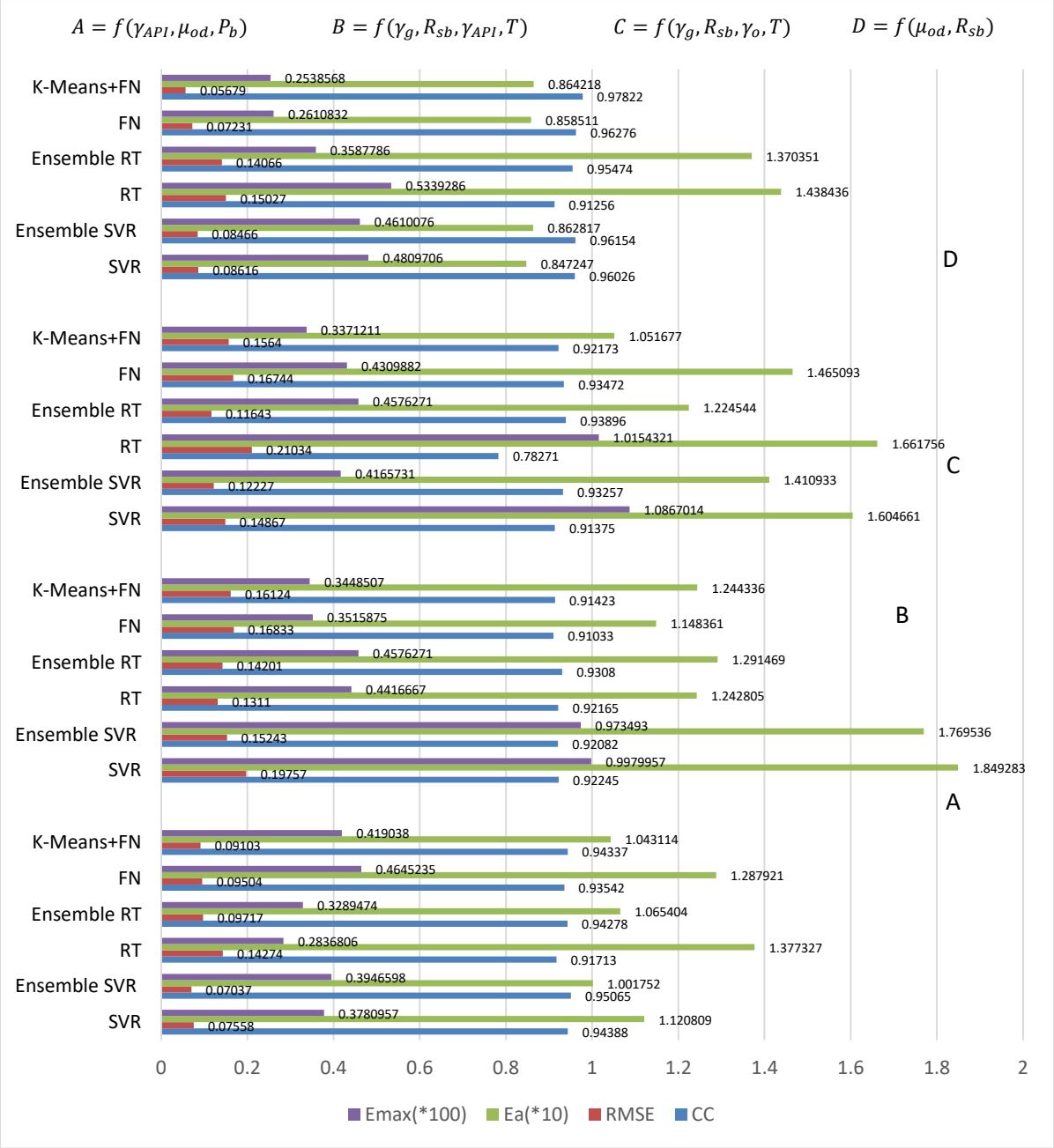


Figure 5.30. Performance of ML techniques for μ_{ob} prediction from data set F

The performance of the best empirical correlation, Al-Khafaji et. al. (1987), in this experimentation with data set F trails the performances of SVR, ensemble SVR, standalone FN and hybrid K-Means+FN with functional form $f(\mu_{od}, R_{sb})$, while the performance of Al-Khafaji et. al. (1987) is better than some ML techniques especially for the functional forms $f(\gamma_g, R_{sb}, \gamma_{API}, T)$ and $f(\gamma_g, R_{sb}, \gamma_o, T)$. It is noted that Al-Khafaji et. al. (1987) uses the same functional form in the main instances where we have four ML techniques performing better than it.

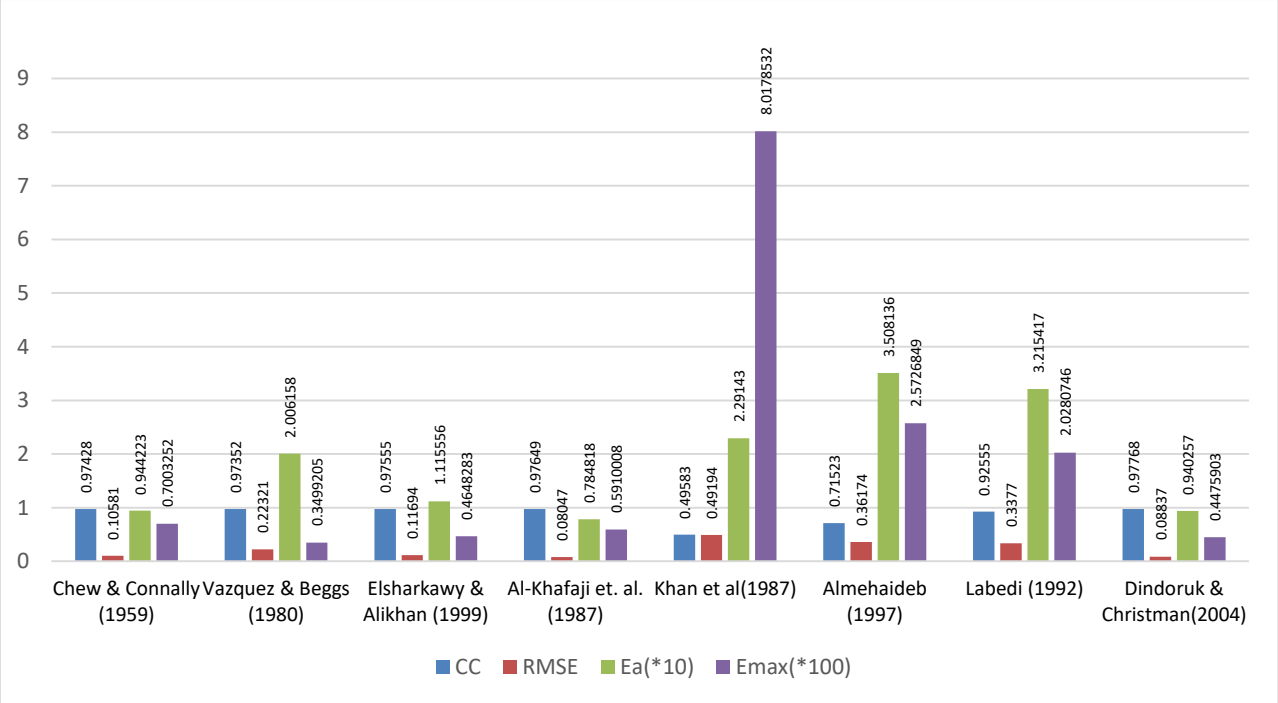


Figure 5.31. Performance of μ_{ob} Empirical Correlations on data set G

Figure 5.31 shows the performances of the empirical correlations on μ_{ob} data set G. Among these methods, Al-Khafaji et. al. (1987) gives the best result with lowest error values of *RMSE* and E_a. The next best performing empirical correlation result is from Dindoruk & Christman(2004). Khan et. al. (1987) gives a very poor result with a very high E_{max}.

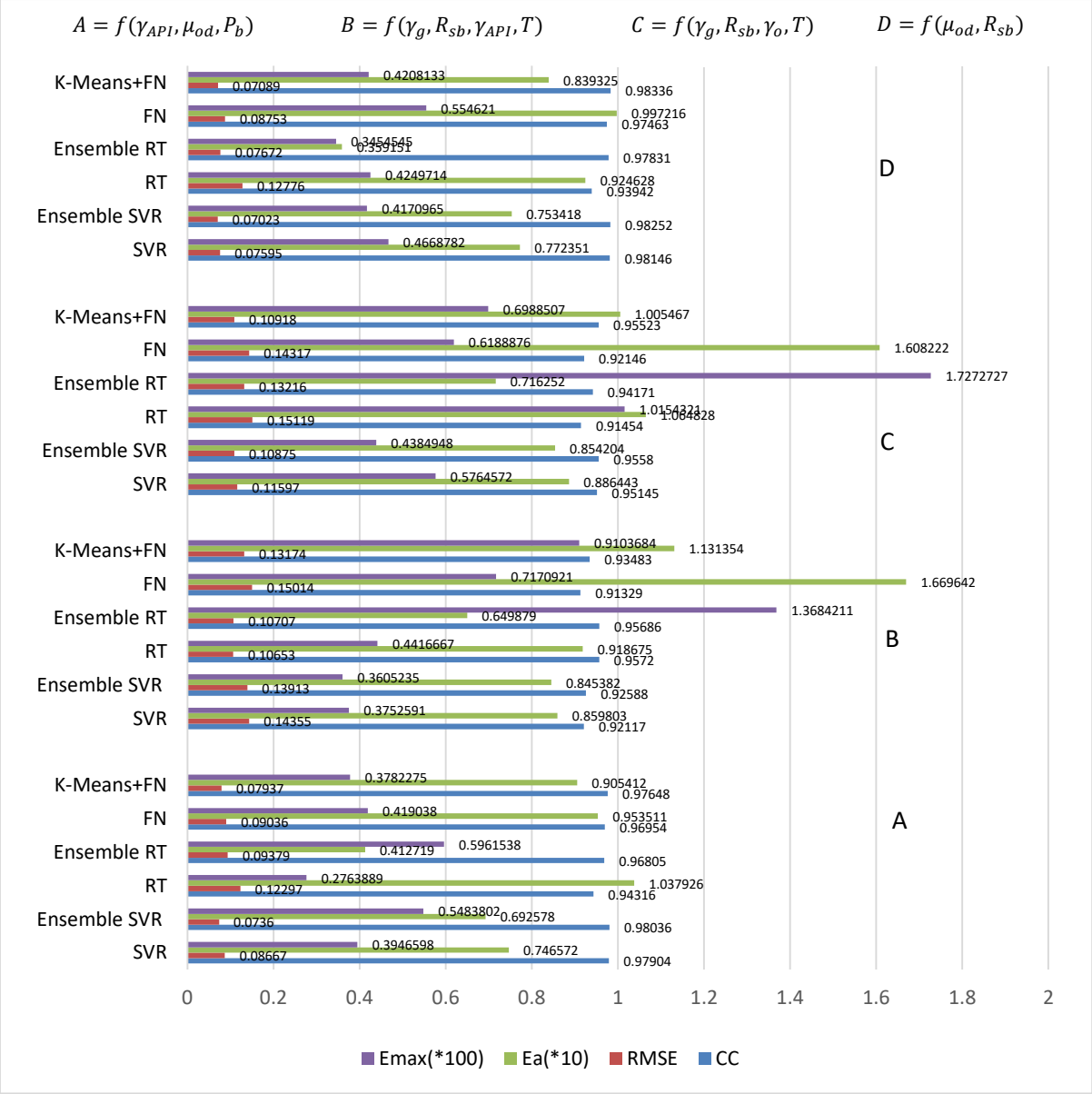


Figure 5.32. Performance of ML techniques for μ_{ob} prediction from data set G

The performances of the ML techniques for μ_{ob} prediction with data set G are shown in Figure 5.32 for different functional forms. The overall best performance is between ensemble SVR with minimum *RMSE* and competitive values of CC, E_a and E_{max}, and ensemble RT with lowest E_a and competitive values of *CC*, *RMSE* and E_{max}, both with the functional form $f(\mu_{od}, R_{sb})$. For this same functional form, the hybrid model K-Means+FN also shows a competitive performance with the overall highest *CC*. Comparable results for to this functional form are also shown by the ML techniques for the functional form $f(\gamma_{API}, \mu_{od}, P_b)$. The results for the other two functional forms ($f(\gamma_g, R_{sb}, \gamma_{API}, T)$ and $f(\gamma_g, R_{sb}, \gamma_o, T)$) are not comparably good.

Though the results of the best two ML techniques on the data set G is better than the best performing empirical correlation, Al-Khafaji et. al. (1987), some empirical correlations give competitive results with some ML techniques.

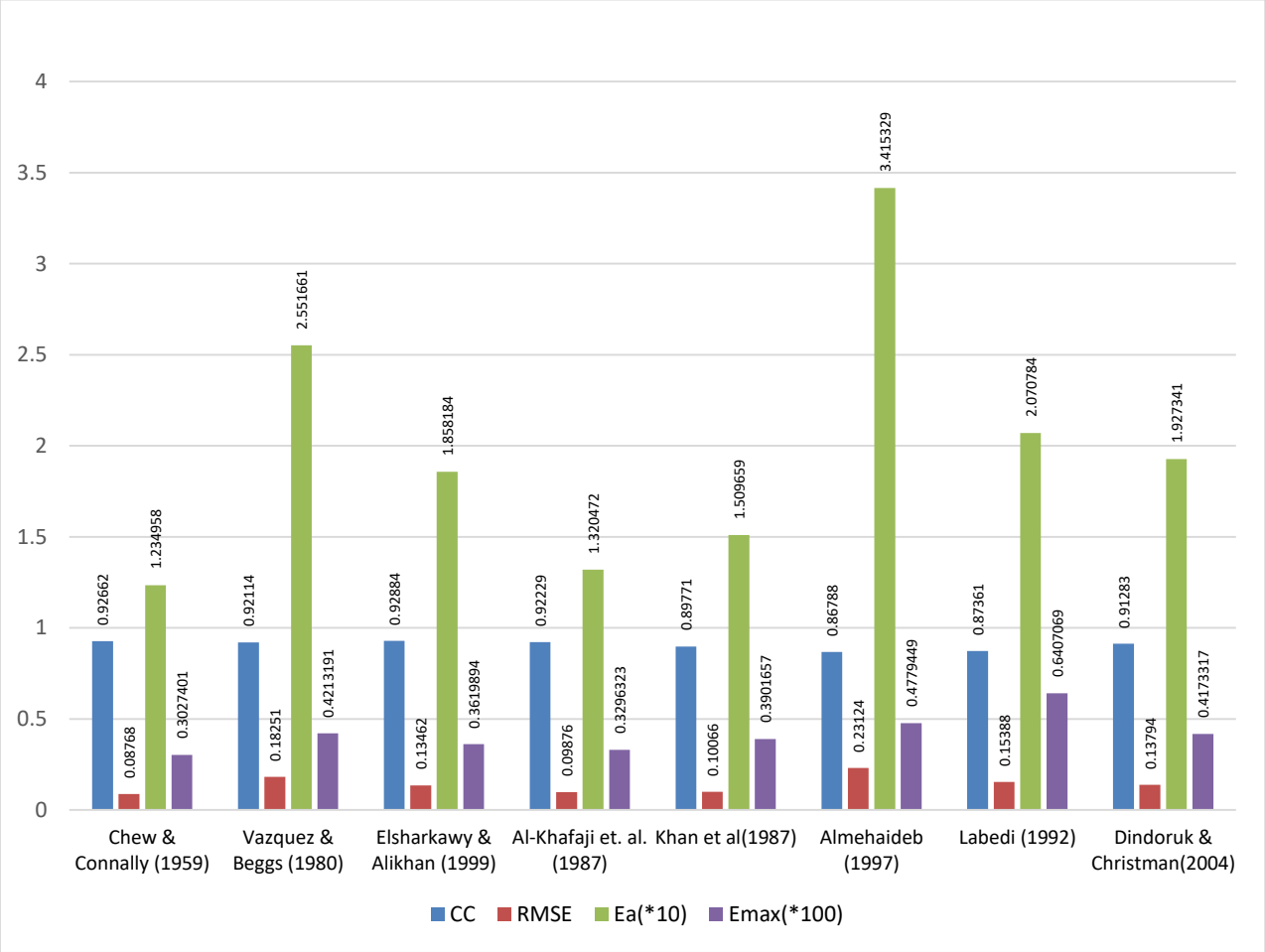


Figure 5.33. Performance of μ_{ob} Empirical Correlations on data set H

Figure 5.33 shows the performances of empirical correlations for μ_{ob} prediction from data set H. Among these correlations, Chew & Connally (1959) gives the best result with minimum RMSE, E_a and E_{max}. It will be noted that the performance of Al-Khafaji et. al. (1987) which was the best among the empirical correlation on μ_{ob} data set G has reduced significantly.

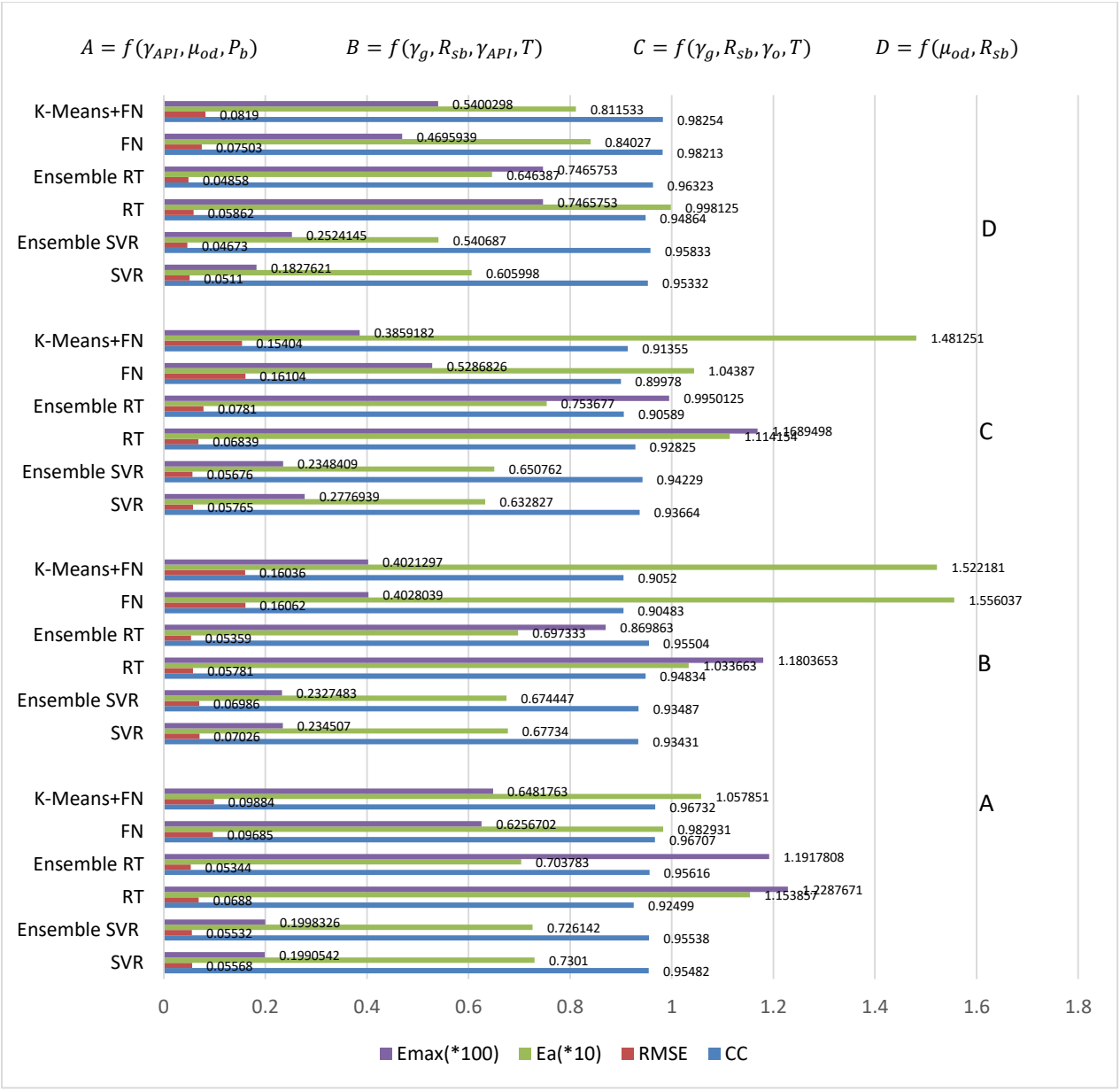


Figure 5.34. Performance of ML techniques for μ_{ob} data set H

The performances of the ML techniques for the μ_{ob} prediction from data set H is shown in Figure 5.34. The ensemble SVR for functional form $f(\mu_{od}, R_{sb})$ has the smallest RMSE and E_a (5.40687). Considering the function form $f(\mu_{od}, R_{sb})$, which gives the best results for data sets F and G, both standalone FN and K-Means+FN have competitive results. This trend is very similar to the case of μ_{od} prediction from data set H. Since data set H is smaller than data sets F and G, then its clusters will have fewer data points and the associated FN for each cluster might have not properly captured the uncertainties in the data effectively. This is one of the shortcomings of the hybrid system though it has proved effective on some data sets with better performance than the standalone FN.

Comparing the empirical correlations with the ML techniques, the best empirical correlation, Chew & Connally (1959), trails the best ML technique, ensemble SVR with functional form $f(\gamma_g, R_{sb}, \gamma_{API}, T)$. Even, many other ML techniques outperform all the empirical correlations on this data set, notably, all ML techniques with the functional form $f(\mu_{od}, R_{sb})$. Hence, the functional form $f(\mu_{od}, R_{sb})$ is likely more probable to model μ_{ob} prediction most consistently with good reliability and generalisation when any of the developed ML techniques is considered to be used.

5.5.2. Trend and Comparative Analysis of Saturated Oil Viscosity Prediction

The overall summary of the performances of ML techniques for μ_{ob} is shown in Table 5.12. For both data sets F and G, K-Means+FN gives the best result with the functional form $f(\mu_{od}, R_{sb})$ while the standalone FN is competitive with it for the data set H with the same functional form. From the statistics of the data sets in the appendix, it can be observed that data set H may contain some volatile oil data as the maximum R_s is 2405.7 scf/STB. Typically, as stated in chapter 2, black oil has R_s of 2000 scf/STB or less. K-Means+FN was developed with data set F which has maximum R_s of 1334 scf/STB. Hence the FNs for the clusters might have not adequately learnt the uncertainties in the data set H compared to a standalone FN.

Ensemble SVR and RT outperform their respective standalone methods across all the data sets F, G and H. It can also be observed that the performances of the correlations on data set H dropped when compared to data sets F and G.

Table 5.12. Summary Comparison of ML Technique for μ_{ob} Prediction

	Data Sets		
Methods	F	G	H
SVR vs Ensemble SVR	Ensemble SVR $[f(\mu_{od}, R_{sb})]$	Ensemble SVR $[f(\mu_{od}, R_{sb})]$	Ensemble SVR $[f(\mu_{od}, R_{sb})]$
RT vs Ensemble RT	Ensemble RT $[f(\mu_{od}, R_{sb})]$	Ensemble RT $[f(\gamma_{API}, \mu_{od}, P_b)]$	Ensemble RT $[f(\mu_{od}, R_{sb})]$
FN vs K-Means+FN	K-Means+FN $[f(\mu_{od}, R_{sb})]$	K-Means+FN $[f(\mu_{od}, R_{sb})]$	FN $[f(\mu_{od}, R_{sb})]$

Tables 5.13 and 5.14 summarise the statistical measures for μ_{ob} prediction from empirical correlations and ML models respectively, showing the minimum and maximum values of these measures. To have an overview of the overall performances across all the data sets (F, G and H),

the average of the statistical measures are shown in Figures 6.35 and 6.36 for empirical correlations and ML models respectively.

Table 5.13. Summary of Statistical Measures for μ_{ob} Empirical Correlations

	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
Chew & Connally (1959)	0.9266 / 0.9743	0.088 / 0.106	9.442 / 12.350	30.274 / 70.803
Vazquez & Beggs (1980)	0.9211 / 0.9735	0.183 / 0.223	20.062 / 25.517	34.992 / 42.132
Elsharkawy & Alikhan (1999)	0.9288 / 0.9756	0.117 / 0.135	11.156 / 18.582	36.199 / 46.483
Al-Khafaji et. al. (1987)	0.9223 / 0.9765	0.080 / 0.099	7.848 / 13.205	32.963 / 60.100
Khan et al(1987)	0.4958 / 0.8977	0.101 / 0.492	15.097 / 22.914	39.017 / 906.779
Almehaideb (1997)	0.7152 / 0.8679	0.231 / 0.362	34.153 / 35.081	47.794 / 287.237
Labedi (1992)	0.8736 / 0.9257	0.154 / 0.338	20.708 / 32.154	64.071 / 252.807
Dindoruk & Christman(2004)	0.9128 / 0.9777	0.088 / 0.138	9.403 / 19.273	41.733 / 44.759

Table 5.14. Summary of Statistical Measures for Testing u_{oa} Prediction with ML techniques

$f(\gamma_{API}, \mu_{od}, P_b)$	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
SVR	0.9439 / 0.9790	0.056 / 0.087	7.301 / 11.208	19.905 / 39.466
Ensemble SVR	0.9507 / 0.9804	0.055 / 0.074	6.926 / 10.018	19.983 / 54.838
RT	0.9171 / 0.9432	0.069 / 0.143	10.379 / 13.773	27.639 / 122.877
Ensemble RT	0.9428 / 0.9681	0.053 / 0.097	4.127 / 10.654	32.895 / 119.178
FN	0.9354 / 0.9695	0.090 / 0.097	9.535 / 12.879	41.904 / 62.567
K-Means+FN	0.9434 / 0.9765	0.079 / 0.099	9.054 / 10.579	37.823 / 64.818
$f(\gamma_g, R_{sb}, \gamma_{API}, T)$				
SVR	0.9212 / 0.9343	0.070 / 0.198	6.773 / 18.493	23.451 / 99.800
Ensemble SVR	0.9208 / 0.9349	0.070 / 0.152	6.744 / 17.695	23.275 / 97.349
RT	0.9217 / 0.9572	0.058 / 0.131	9.187 / 12.428	44.167 / 118.037
Ensemble RT	0.9308 / 0.9569	0.054 / 0.142	6.499 / 12.915	45.763 / 136.842
FN	0.9048 / 0.9133	0.150 / 0.168	11.484 / 16.696	35.159 / 71.709
K-Means+FN	0.9052 / 0.9348	0.132 / 0.161	11.314 / 15.222	34.485 / 91.037
$f(\gamma_g, R_{sb}, \gamma_o, T)$				
SVR	0.9138 / 0.9515	0.058 / 0.149	6.328 / 16.047	27.769 / 108.670
Ensemble SVR	0.9326 / 0.9558	0.057 / 0.122	6.508 / 14.109	23.484 / 43.849
RT	0.7827 / 0.9283	0.068 / 0.210	10.648 / 16.618	101.543 / 116.895
Ensemble RT	0.9059 / 0.9417	0.078 / 0.132	7.163 / 12.245	45.763 / 172.727

FN	0.8998 / 0.9347	0.143 / 0.167	10.439 / 16.082	43.099 / 61.889
K-Means+FN	0.9136 / 0.9552	0.109 / 0.156	10.055 / 14.813	33.712 / 69.885
$f(\mu_{od}, R_{sb})$				
SVR	0.9533 / 0.9815	0.051 / 0.086	6.060 / 8.472	18.276 / 48.097
Ensemble SVR	0.9583 / 0.9825	0.047 / 0.085	5.407 / 8.628	25.241 / 46.101
RT	0.9126 / 0.9486	0.059 / 0.150	9.246 / 14.384	42.497 / 74.658
Ensemble RT	0.9547 / 0.9783	0.049 / 0.141	3.592 / 13.704	34.545 / 74.658
FN	0.9628 / 0.9821	0.072 / 0.088	8.403 / 9.972	26.108 / 55.462
K-Means+FN	0.9782 / 0.9834	0.057 / 0.082	8.393 / 9.115	25.386 / 54.003

Since most ML techniques in Table 6.14 have their best results with the functional form $f(\mu_{od}, R_{sb})$, this group will be used for comparison with the results of the empirical correlations. The following can be observed from these tables.

- All ML models from this group except RT have better minimum and maximum CC than all the empirical correlations. The average measures in Figures 5.35 and 5.36 also corroborates the overall results for the CC as only the RT trails some of the empirical correlations while all other ML models have higher average performances with respect to this statistical measure.
- For the minimum and maximum RMSE and E_a, four ML models (SVR, ensemble SVR, FN and K-Means+FN) clearly outperform all the empirical correlations. Again, the average measures (Figures 5.35 and 5.36) also corroborate these results as these four ML models have lower average RMSE and E_a than the empirical correlations.
- For the E_{max}, the average, minimum and maximum values of this statistical measure for some empirical correlations are very competitive with the four leading ML models. This implies that these methods have similar over-predictions for some data points but not necessarily the same ones.
- Given that the four leading ML techniques have clearer better performances with respect to CC, RMSE and E_a, their overall and average performances are better than the empirical correlations.

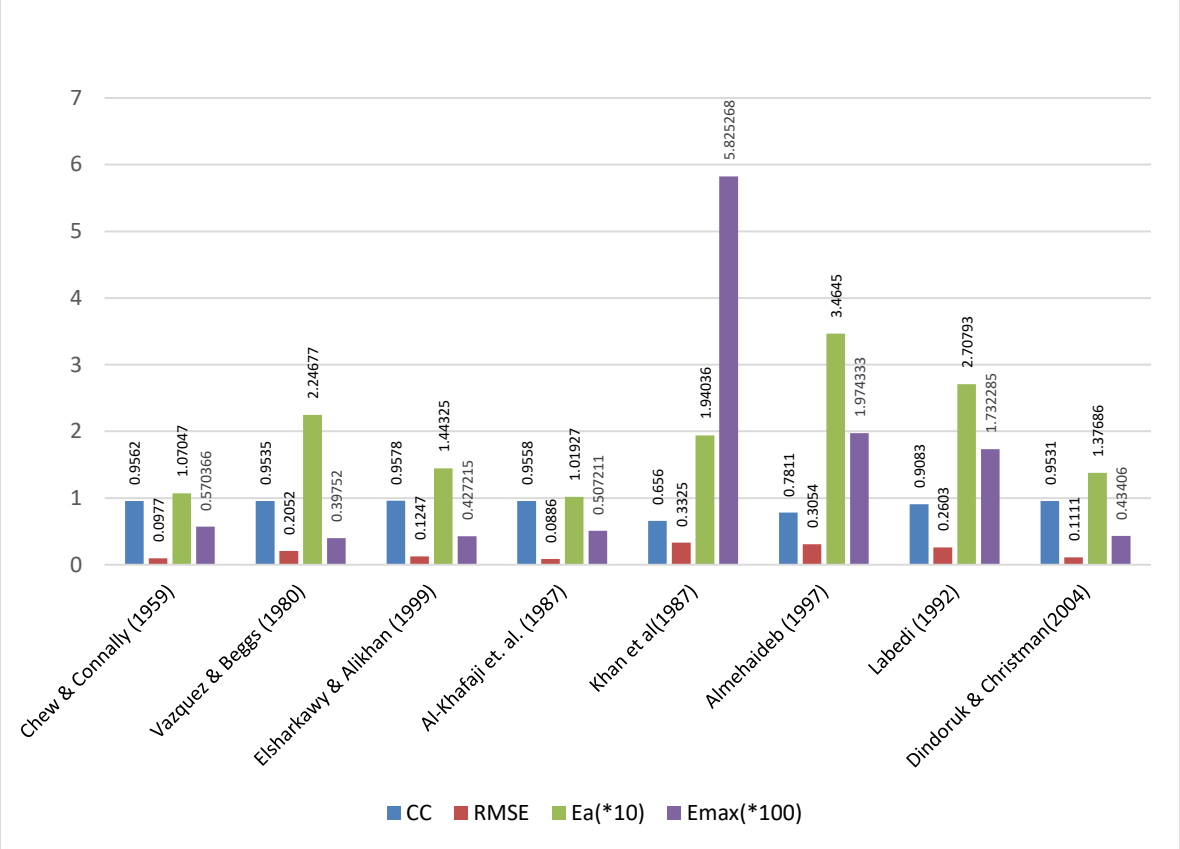


Figure 5.35. Average of Statistical Measures for μ_{ob} Empirical Correlations

5.6. Prediction of Undersaturated Oil Viscosity

Three different functional forms for μ_{oa} have been examined in this study. These are as follow.

$$\mu_{oa} = f(\mu_{ob}, P_b, P)$$

$$\mu_{oa} = f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$$

$$\mu_{oa} = f(\mu_{ob}, \mu_{od} P_b, P)$$

The experimental work in this section attempts to evaluate the impact of the μ_{od} and/or γ_{API} in modelling μ_{oa}. The evaluation examine if the additional parameter improves or can improve the prediction of the target output, μ_{oa}.

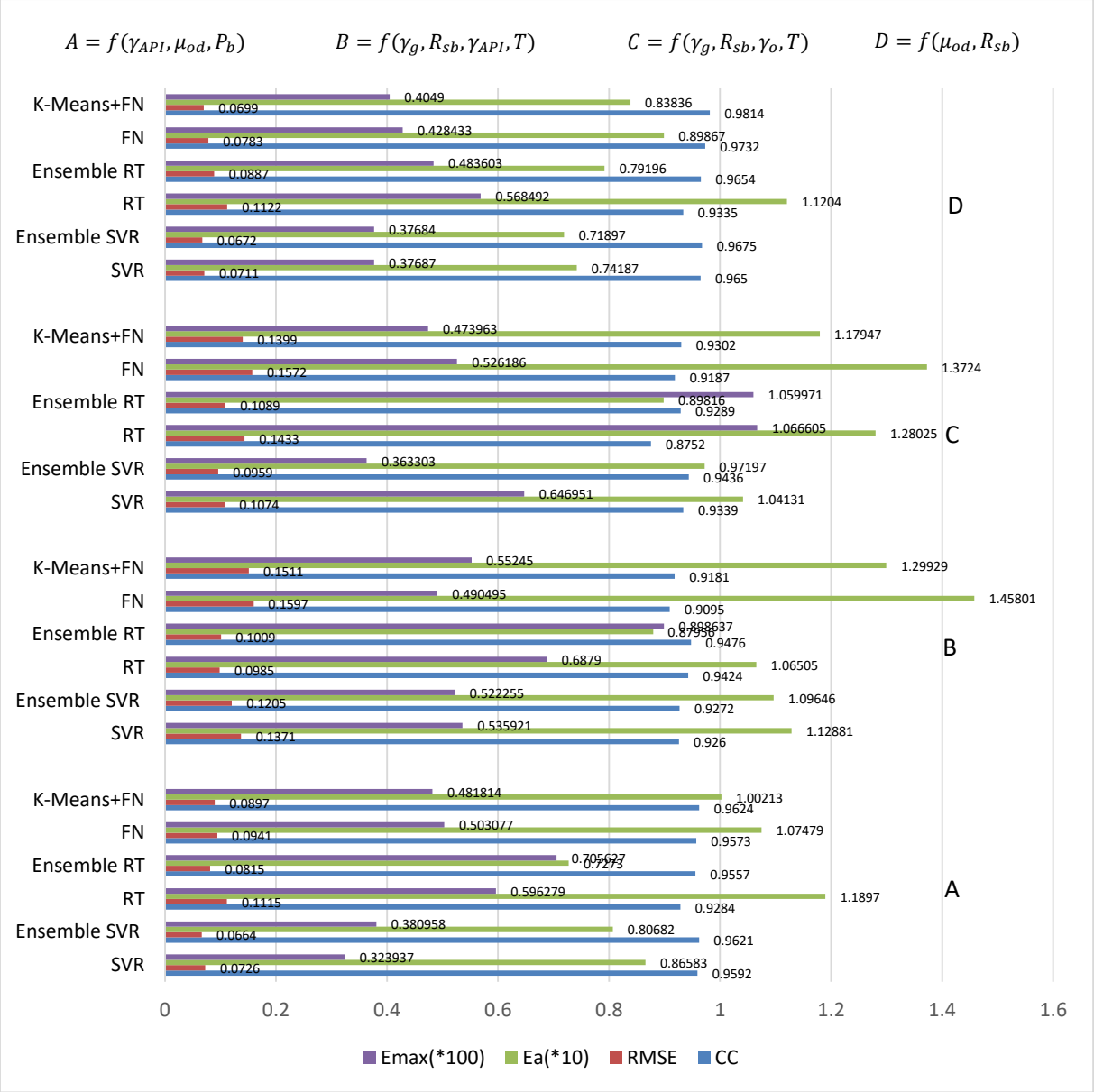


Figure 5.36. Average of Statistical Measures for Testing u_{ob} Prediction with ML techniques

Table 5.15. Categorisation of the Evaluated μ_{oa} Correlations Based on their Functional Forms

Correlation Methods	Functional Form
Beal (1946); Vazquez & Beggs (1980)	μ_{ob}, P_b, P
Labedi (1992)	$\mu_{ob}, P_b, P, \mu_{od}, \gamma_{API}$
Elsharkawy and Alikhan (1999)	$\mu_{ob}, P_b, P, \mu_{od}$

5.6.1. Results and Error Analysis for Undersaturated Oil Viscosity

The results for the prediction of μ_{oa} are shown in Figures 5.37 to 5.42. Different models have been developed for the six ML techniques using the three stated functional forms. Results of the developed ML models have been compared with selected empirical correlations.

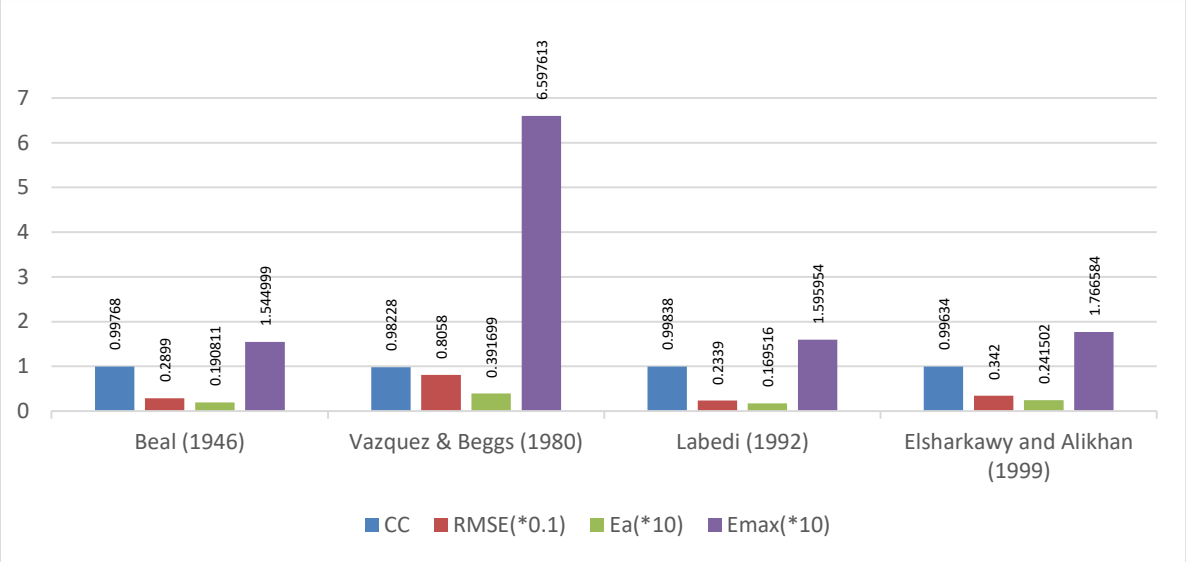


Figure 5.37. Performance of μ_{oa} Empirical Correlations on data set I

For the viscosity data set I, Labedi (1992) has the best performance among the compared empirical correlations with highest *CC* and lowest *RMSE* and E_a. Beal (1946) shows a competitive result with the lowest E_{max}.

On viscosity data set I, ensemble SVR with functional forms $f(\mu_{ob}, \mu_{od}P_b, P)$ and $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ have the best performances where the former has competitive value of *CC* and error values of *RMSE*, E_a and E_{max}. The latter has the highest *CC* and error values of *RMSE*, E_a and E_{max}. The former ensemble SVR has the lowest *RMSE* and E_{max}, while the latter has the highest CC and lowest E_a. The performance of the ensemble RT is better than the standalone RT, virtually in all cases. The hybrid K-Means+FN also shows improved performance than the standalone FN in most cases with respect to the four metric values (*CC, RMSE*, E_a and E_{max}).

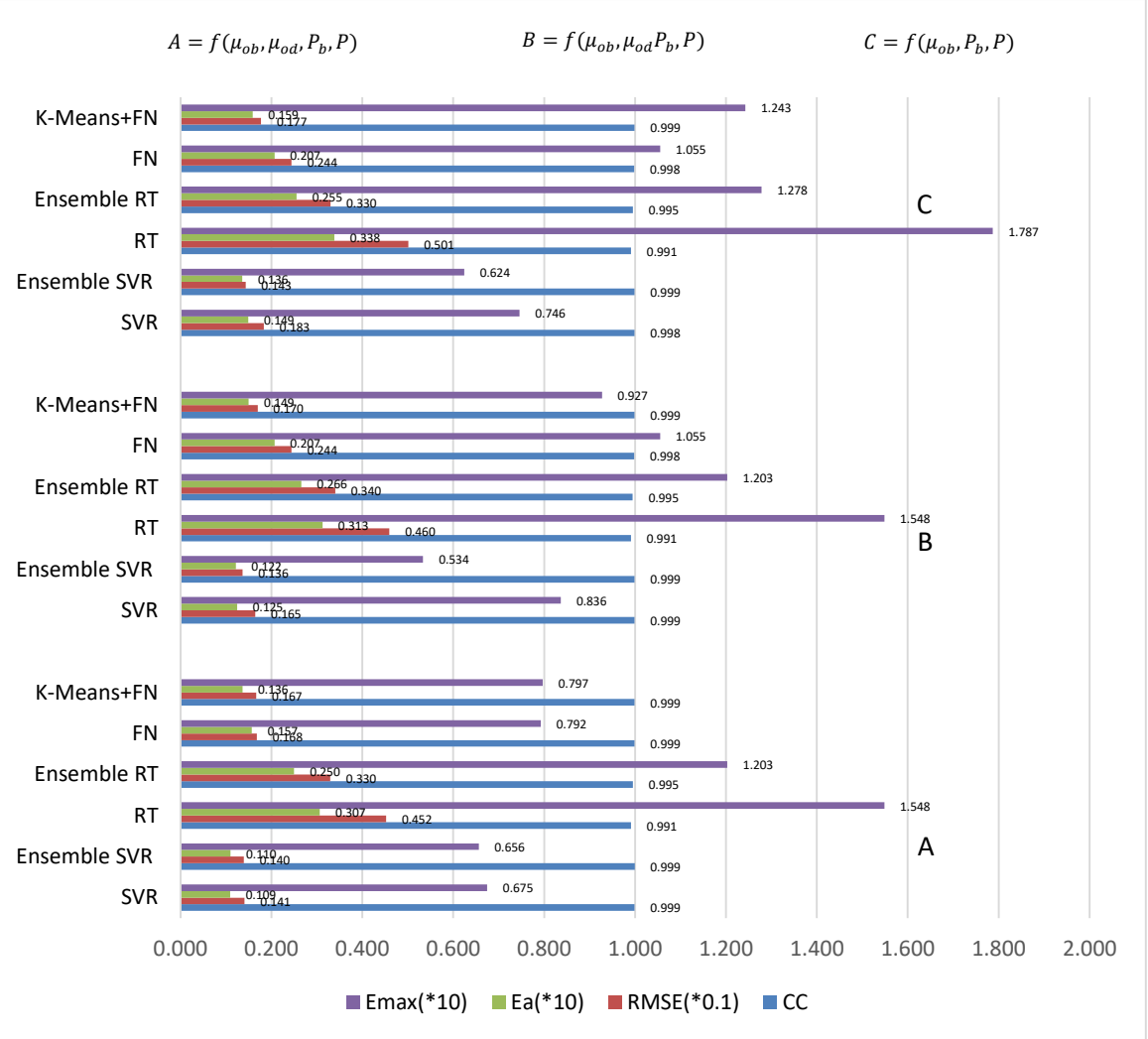


Figure 5.38. Performance of ML techniques for μ_{oa} data set I

The best performing empirical correlation, Labedi (1992), trails some ML techniques in performance, such as SVR, ensemble SVR , FN and hybrid K-Means+FN with functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$, SVR, ensemble SVR and K-Means+FN with functional forms $f(\mu_{ob}, \mu_{od}P_b, P)$ and $f(\mu_{ob}, P_b, P)$. Though the performance of the ensemble RT is better than some of the evaluated empirical correlations, it trails the performance of Labedi (1992).

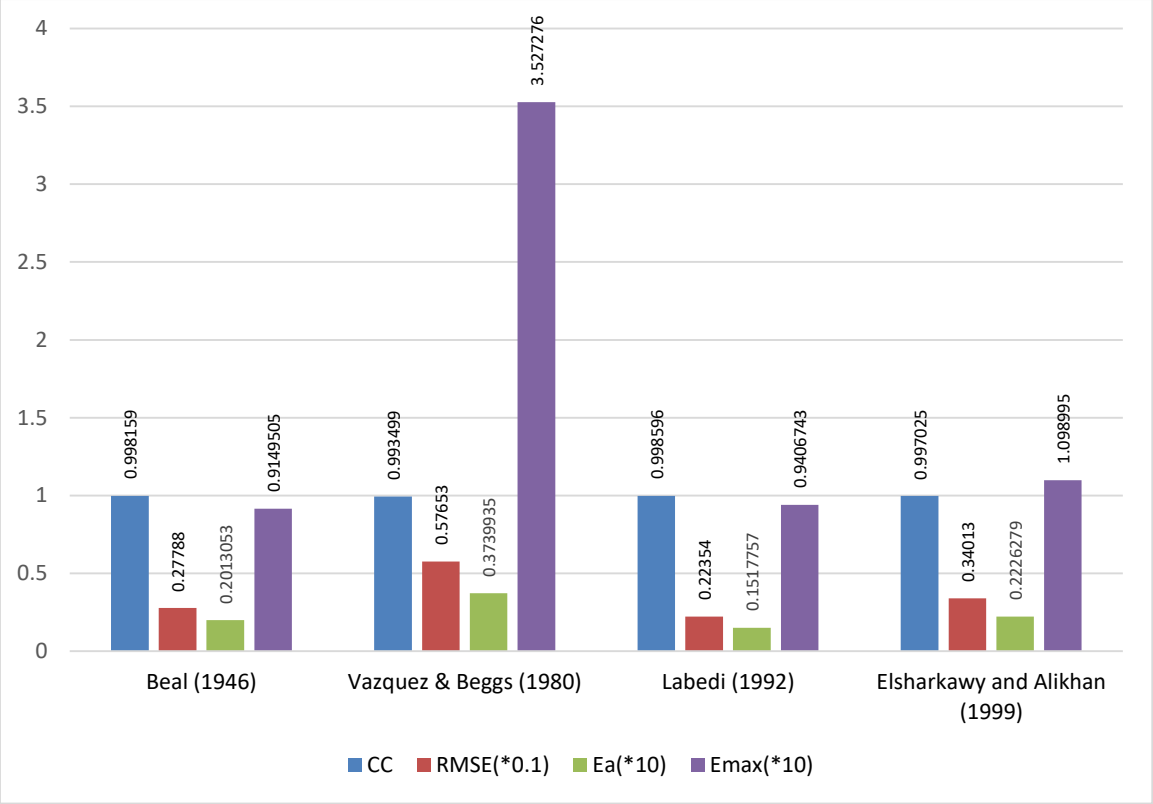


Figure 5.39. Performance of μ_{oa} Empirical Correlations on data set J

In Figure 5.39, the empirical correlation with the best performance is Labedi (1992) with the highest *CC* and lowest error values for *RMSE* and E_a. The correlation of Beal (1946) gives minimum E_{max} and its overall performance is very competitive with that of Labedi (1992).

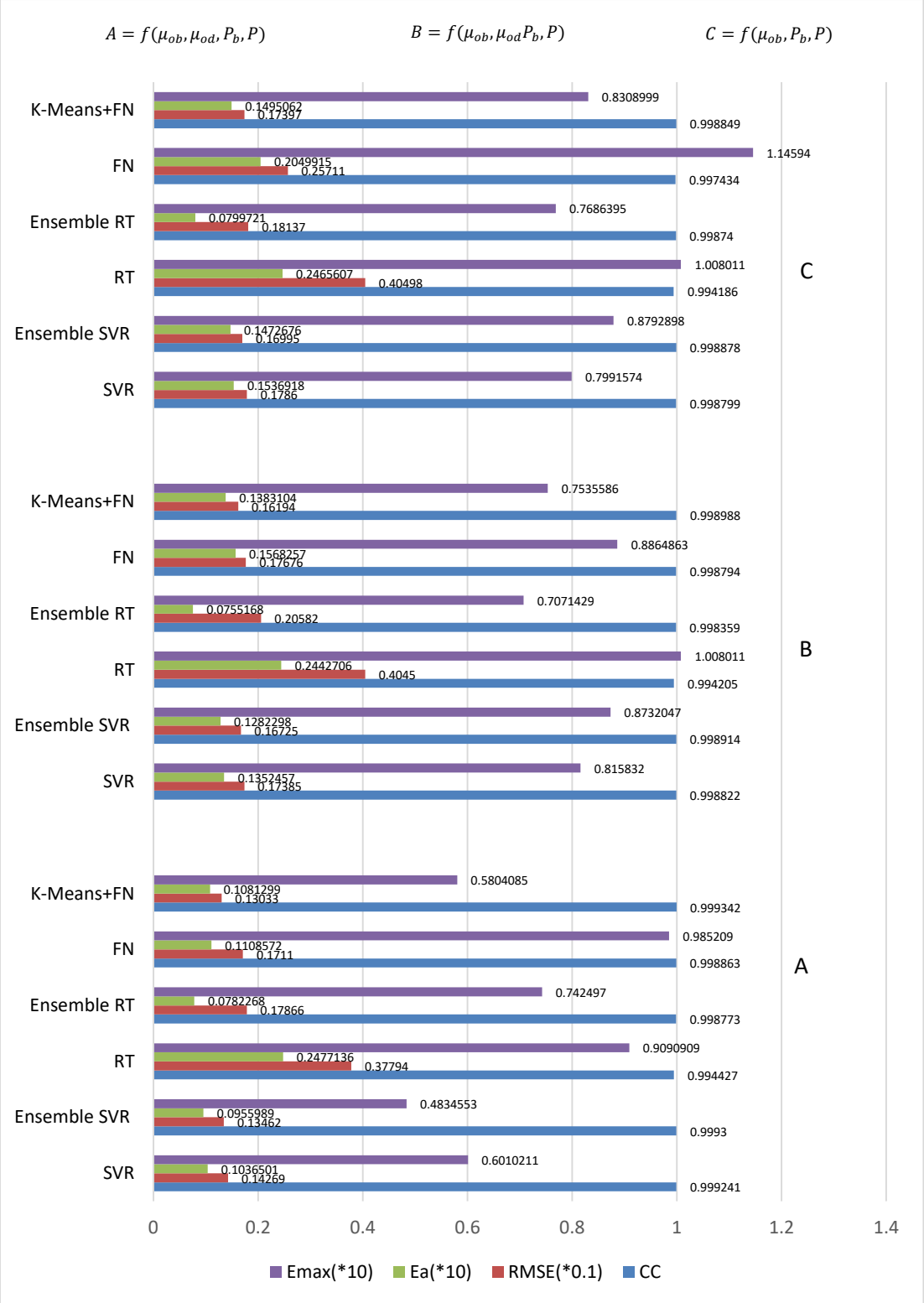


Figure 5.40. Performance of ML techniques for μ_{oa} data set J

Figure 5.40 gives the experimental results of ML techniques on data set J. In this case, the ensemble SVR and hybrid K-Means+FN with the functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ have the best performance where the former has the lowest values of E_{max} while the latter has the highest *CC*

and lowest *RMSE*. The ensemble RT has competitive result in this category with the lowest E_a but its other metric values are not as good as the aforementioned best two techniques for this data set.

It is noted that the performances of Beal (1946) and Labedi (1992) are better than the results of some ML techniques, notably, standalone FN and RTs for functional forms $f(\mu_{ob}, \mu_{od} P_b, P)$ and $f(\mu_{ob}, P_b, P)$, but they trail the results of ensemble SVR, ensemble RT and hybrid K-Means+FN for all the functional forms.

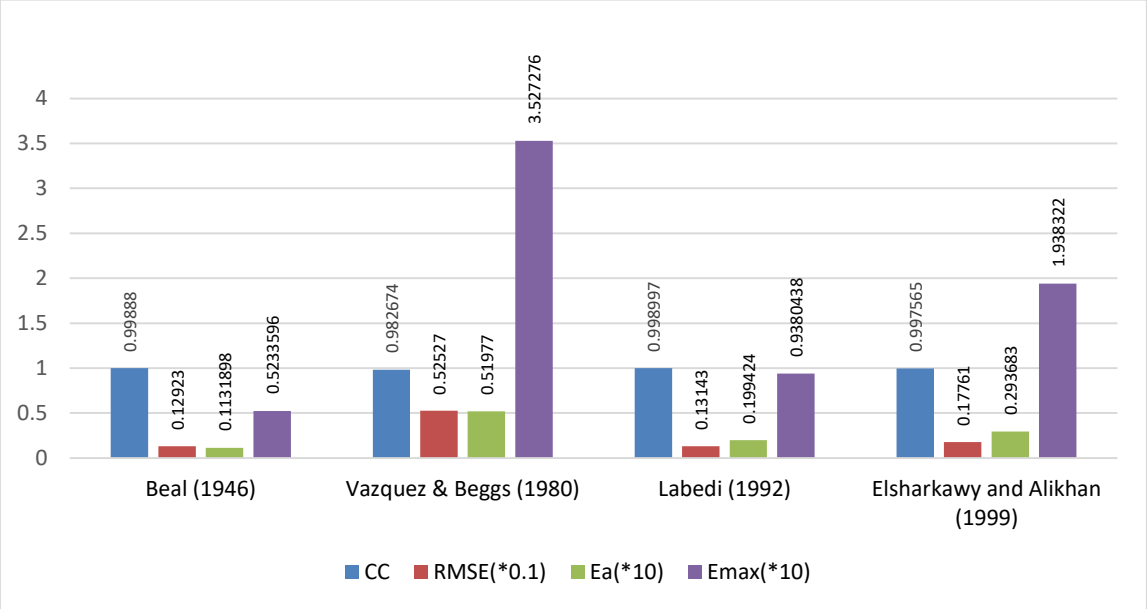


Figure 5.41. Performance of μ_{oa} Empirical Correlations on data set K

The performance of empirical correlations for μ_{oa} on data set K is given in Figure 5.41. Beal (1946) gives the best result across the three error metric measures with minimum values of RMSE, E_a and E_{max} while Labedi (1992) has the highest *CC* which is slightly higher than that of Beal (1946). Vazquez & Beggs (1980) has the poorest performance with the highest error values for RMSE, E_a and E_{max}.

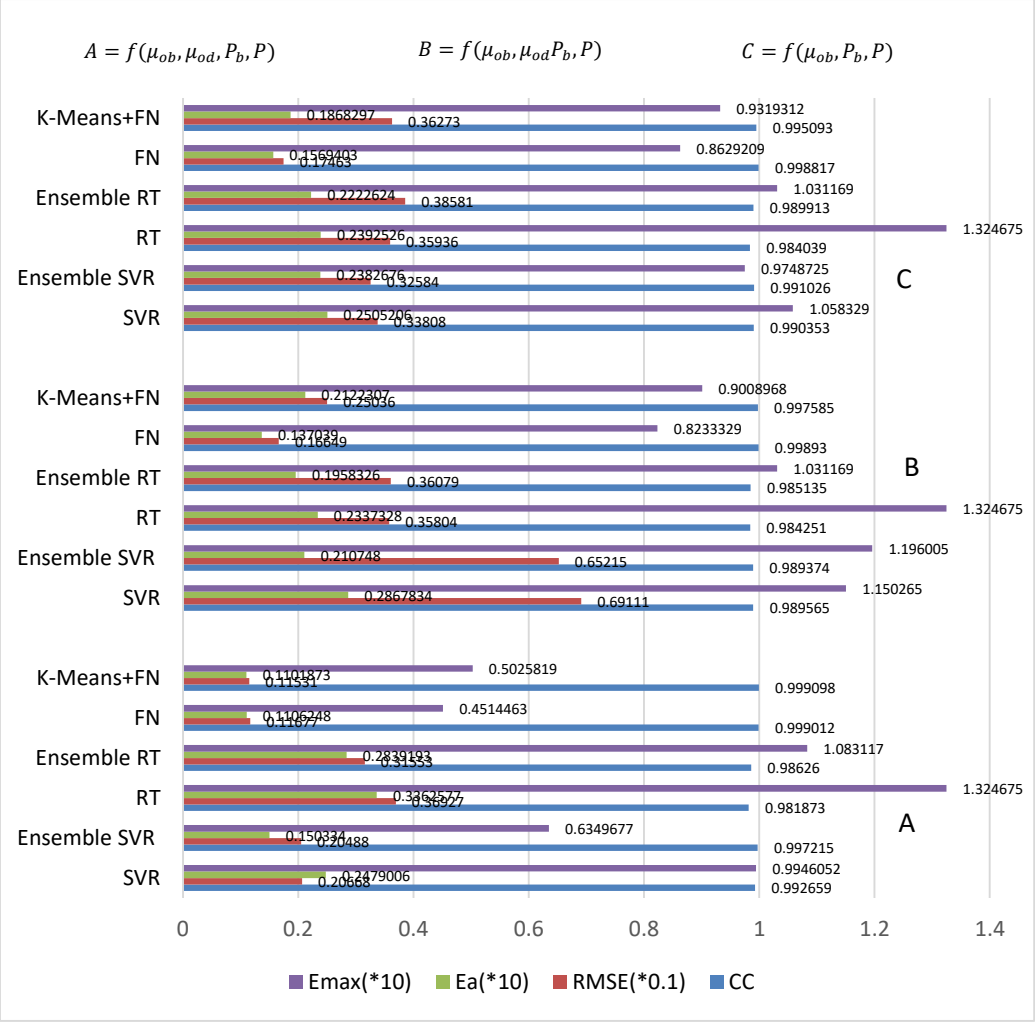


Figure 5.42. Performance of ML techniques for μ_{oa} data set K

Figure 5.42 shows the performance of the ML techniques for μ_{oa} prediction on data set K. Among the ML techniques, the standalone FN and its hybrid K-Means+FN with the functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ have the best performance with the former having lowest E_a and E_{max} while the latter has the highest CC and lowest RMSE. The performance of the ensemble SVR for this functional form seems consistent with the previous data sets of I and J though is lower than the performance of FN and K-Means+FN on this data set. The performances of all the ML techniques on the data set K for the functional forms $f(\mu_{ob}, \mu_{od} P_b, P)$ and $f(\mu_{ob}, P_b, P)$ are poorer than the previous data sets: I and J. Possibly, these functional forms are not true representation of μ_{oa} or do not capture the uncertainties in it as much as $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$.

The correlation of Beal (1946) shows competitive performance with FN and K-Means+FN with the functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ and better performance than many ML techniques on this

data set K. However, the performance its previous performances have been lower than some ML techniques especially ensemble SVR with the functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$. Beal (1946) has shown a much wider deviations in performances across all the three data sets for μ_{oa} prediction. Subsequent section sheds more light on this.

5.6.2. Trend and Comparative Analysis of Undersaturated Oil Viscosity Prediction

More input correlating variables tend to have impact in the performances of the ML techniques for the prediction of μ_{oa}. The functional form with the highest number of variables, $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$, gives best results in most cases for each ML technique or in comparative category. Given the results in section 5.6.1, all ML techniques give generally good results and at least one or more of them give better results than the empirical correlations.

As noted in chapter 3, the empirical correlations for μ_{oa} usually perform well. It will be observed that the γ_{API} of all μ_{oa} data sets (I, J and K) have close maximum values though they have varying minimum values. Hence the possibility of very similar trends in the performances of each of the observed correlations and the ML techniques. It is noted that the maximum values of the γ_{API} are within the maximum value for black oil as earlier mentioned.

Table 5.16. Deductive Comparison of ML Technique for μ_{oa} Prediction

	Data Sets		
Methods	I	J	K
SVR vs Ensemble SVR	Ensemble SVR with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ or $f(\mu_{ob}, \mu_{od} P_b, P)$	Ensemble SVR with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$	Ensemble SVR with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$
RT vs Ensemble RT	Ensemble RT with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$	Ensemble RT with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$	Ensemble RT with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ or $f(\mu_{ob}, \mu_{od} P_b, P)$
FN vs K-Means+FN	K-Means+FN with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$	K-Means+FN with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$	Tie between FN and K-Means+FN with $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$

The summary of the minimum and maximum values of the statistical measures for μ_{oa} predictions from ML models of both experiments 1 and 2 and the empirical correlations for data sets I, J and K is shown in Tables 5.17 and 5.18.

Table 5.17. Summary of Statistical Measures for μ_{oa} Empirical Correlations

	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
Beal (1946)	0.9977 / 0.9989	0.013/ 0.029	1.132 / 2.013	5.234 / 15.450
Vazquez & Beggs (1980)	0.9823 / 0.9935	0.053/ 0.081	3.740 / 5.198	35.273 / 65.976
Labedi (1992)	0.9984 / 0.9990	0.013/ 0.023	1.518 / 1.994	9.380 / 15.960
Elsharkawy and Alikhan (1999)	0.9963 / 0.9976	0.018/ 0.034	2.226 / 2.937	10.990 / 19.383

Table 5.18. Summary of Statistical Measures for Testing u_{oa} Prediction with ML techniques

$f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$	CC(min/max)	$RMSE$(min/max)	E_a(min/max)	E_{max}(min/max)
SVR	0.9927 / 0.9992	0.014/ 0.021	1.037 / 2.479	6.010 / 9.946
Ensemble SVR	0.9972 / 0.9993	0.013/ 0.020	0.956 / 1.503	4.835 / 6.562
RT	0.9819 / 0.9944	0.037/ 0.045	2.477 / 3.363	9.091 / 15.484
Ensemble RT	0.9863 / 0.9988	0.018/ 0.033	0.782 / 2.839	7.425 / 12.030
FN	0.9989 / 0.9990	0.012/ 0.017	1.106 / 1.570	4.514 / 9.852
K-Means+FN	0.9990 / 0.9993	0.012/ 0.017	1.081 / 1.362	5.026 / 7.966
$f(\mu_{ob}, \mu_{od} P_b, P)$				
SVR	0.9896 / 0.9988	0.016/ 0.069	1.245 / 2.868	8.158 / 11.503
Ensemble SVR	0.9894 / 0.9989	0.014/ 0.065	1.218 / 2.107	5.336 / 11.960
RT	0.9843 / 0.9942	0.036/ 0.046	2.337 / 3.125	10.080 / 15.484
Ensemble RT	0.9851 / 0.9984	0.021/ 0.036	0.755 / 2.658	7.071 / 12.030
FN	0.9978 / 0.9989	0.017/ 0.024	1.370 / 2.068	8.233 / 10.551
K-Means+FN	0.9976 / 0.9990	0.016/ 0.025	1.383 / 2.122	7.536 / 9.272
$f(\mu_{ob}, P_b, P)$				
SVR	0.9904 / 0.9988	0.018/ 0.034	1.493 / 2.505	7.458 / 10.583
Ensemble SVR	0.9910 / 0.9989	0.014/ 0.033	1.358 / 2.383	6.239 / 9.749
RT	0.9840 / 0.9942	0.036/ 0.050	2.393 / 3.384	10.080 / 17.869
Ensemble RT	0.9899 / 0.9987	0.018/ 0.039	0.800 / 2.551	7.686 / 12.782
FN	0.9974 / 0.9988	0.017/ 0.026	1.569 / 2.068	8.629 / 11.459
K-Means+FN	0.9951 / 0.9989	0.017/ 0.036	1.495 / 1.868	8.309 / 12.425

Tables 5.17 and 5.18 give the summary of minimum and maximum values of the statistical measures from the μ_{oa} prediction results of the comparing empirical correlations and developed ML models respectively. Most ML techniques have their best results with the functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ and thus that can be used for the overall comparison between the ML techniques and the empirical correlations. The following can be inferred from the two tables.

- The results of the minimum and maximum CC for all the empirical correlations and ML techniques are very close and any difference might be insignificant. Likewise, the minimum/maximum *RMSE* for the leading empirical correlations, Beal, (1946) and Labedi, (1992), can be assumed to have insignificant difference to the leading ML techniques (SVR, ensemble SVR, FN and K-Means+FN) for this parameter.
- However, ensemble SVR, FN and K-Means+FN have lower minimum/maximum E_a and E_{max} than all the empirical correlations. This implies that these ML techniques are less likely to over-predict any data point and their average is better.
- Corroborating the points above, Figures 5.43 and 5.45 show that SVR, ensemble SVR, FN and K-Means+FN with functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ have overall better performance, given their lower average RMSE, E_a and E_{max} than all the empirical correlations.

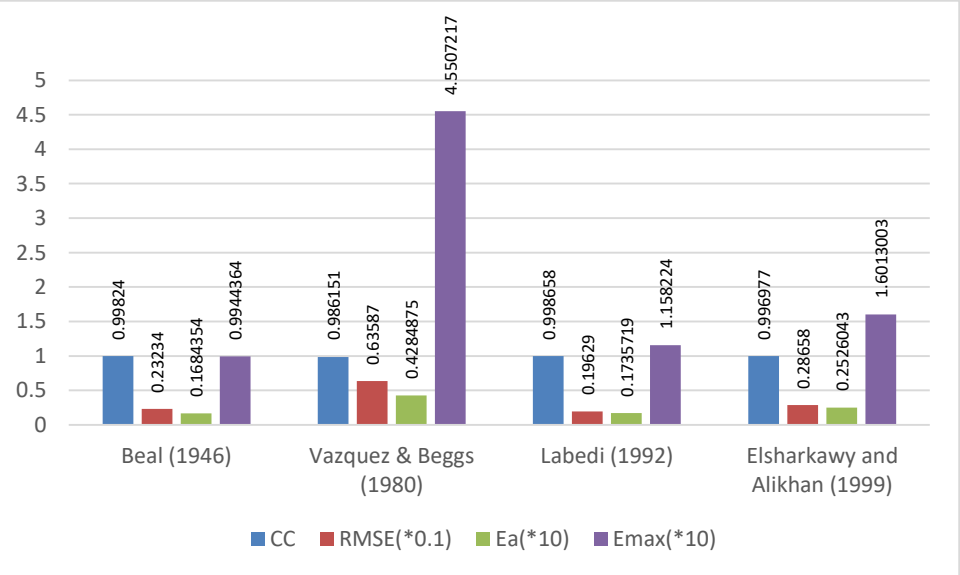


Figure 5.43. Average of Statistical Measures for μ_{oa} Empirical Correlations

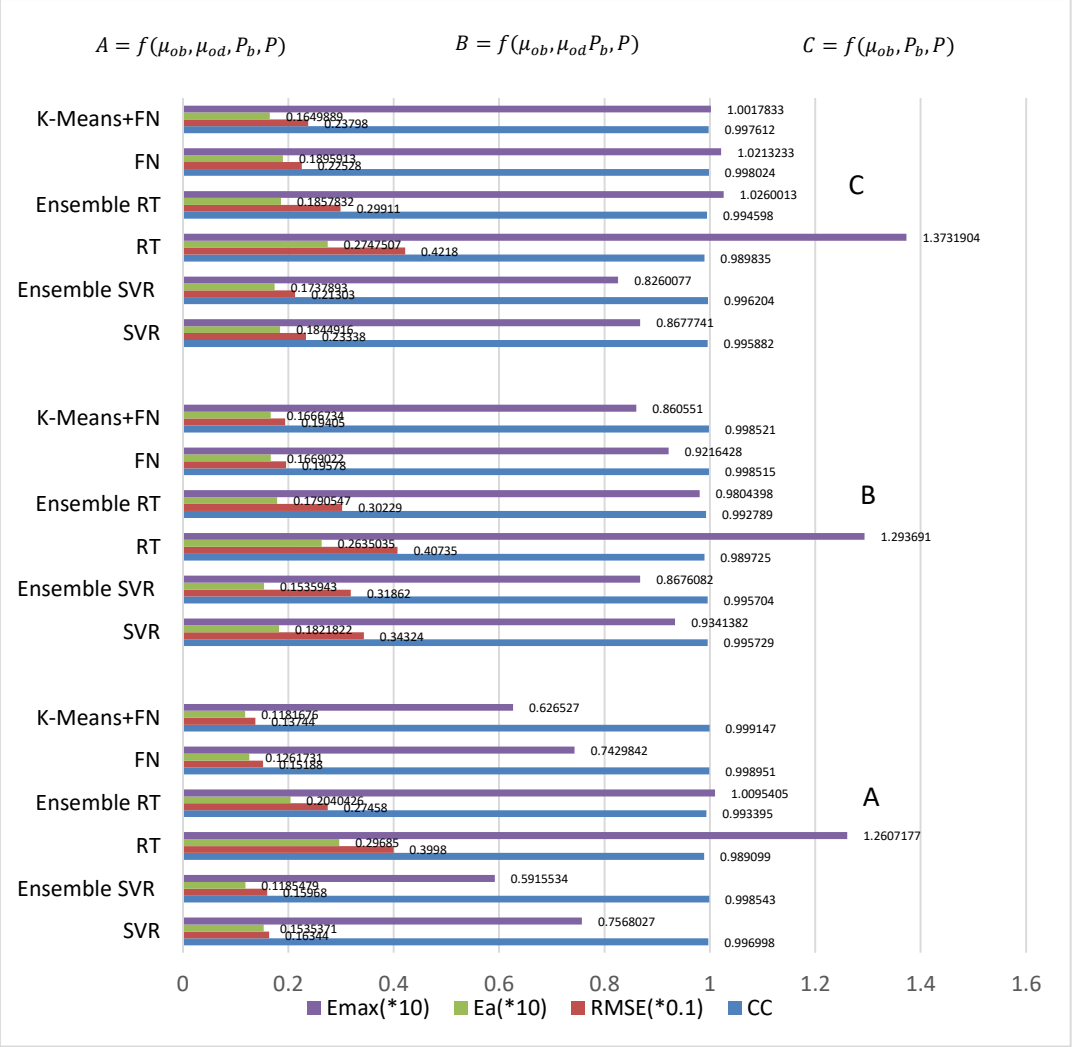


Figure 5.44. Average of Statistical Measures for Testing u_{oa} Prediction with ML techniques

5.7. Impacts of Different Statistical Measures and Graphical Analysis

One of the statistical measures that is commonly reported in the results of PVT analysis and sometimes used to grade the performance of a method is the *SD* (Al-Marhoun, 1988; Petrosky Jr & Farshad, 1998; Olatunji et. al., 2011) . Olatunji et. al., (2011) erroneously rate their methods above comparing methods using SD even when other perfomance measures (CC, E_a and E_{max}) are poorer than the comparing methods or empirical correlations. Emphasis on *SD* rather than other statistical measures such as CC, $RMSE$ and E_a could be misleading as the analysis here will shortly unveil.

Given the following errors (APEs): 30, 34, 36, 37 and 39 for instance, the *SD* and mean are 3.4205 and 35.2 respectively. On the other hand, the *SD* and mean for the errors 4, 9, 12, 14, 23 are 7.0214 and 12.4 respectively. Obviously, using the *SD* to judge the better set of errors is erroneous. Since the aim is to reduce the errors, so the second category of errors is apparently better despite having

higher *SD*. An example is the correlation of performance of Al-Marhoun (1988) on data sets C and D compared with Standing (1947). Cross plots of the results for both methods are shown in Figures 5.45 and 5.46. Occasionally, the best result may also have the minimum *SD*.

The CC can be interpreted in this context to be an indication of how much can the cross plot between the predicted and the actual outputs attain 45^0 slope. For the RMSE, it will be observed that the values are generally larger for P_b compared to other properties investigated in this work. This is because its actual values are generally bigger and up to four digits, compared with other properties which are generally small and rarely have more than a unit. So, the square of the errors also become big.

Also, an interesting statistical measure which is also used in comparing the results of methods in PVT analysis is E_{max}. It is an indication of the maximum over-prediction error obtainable from a method on the given data set. The E_{max} for a method may be comparatively high while its other statistical measures may be better than the comparing method(s), i.e., lesser values. An example is the performance of Petrosky Jr & Farshad (1998) on data set D compared with Jarrahian et al., 2015) or Al-Marhoun (1988) on data sets A against Standing (1947).

Cross plots for ensemble SVR and K-Means+FN from experiment 2 for P_b prediction are shown in Figures 5.48 and 5.49. Looking at their plots for data set E, it will be observed they show better matching than Al-Marhoun (1988) on the same data set despite that they have higher E_{max} is higher. This is because their overall error (RMSE and E_a) are much lower. It should be noted that E_{max} is the error for one of the data points. So, when E_{max} is high, the errors for other data points could be much lower which will be essentially reflected in E_a. Some cross plots for other predicted variables have also been presented.

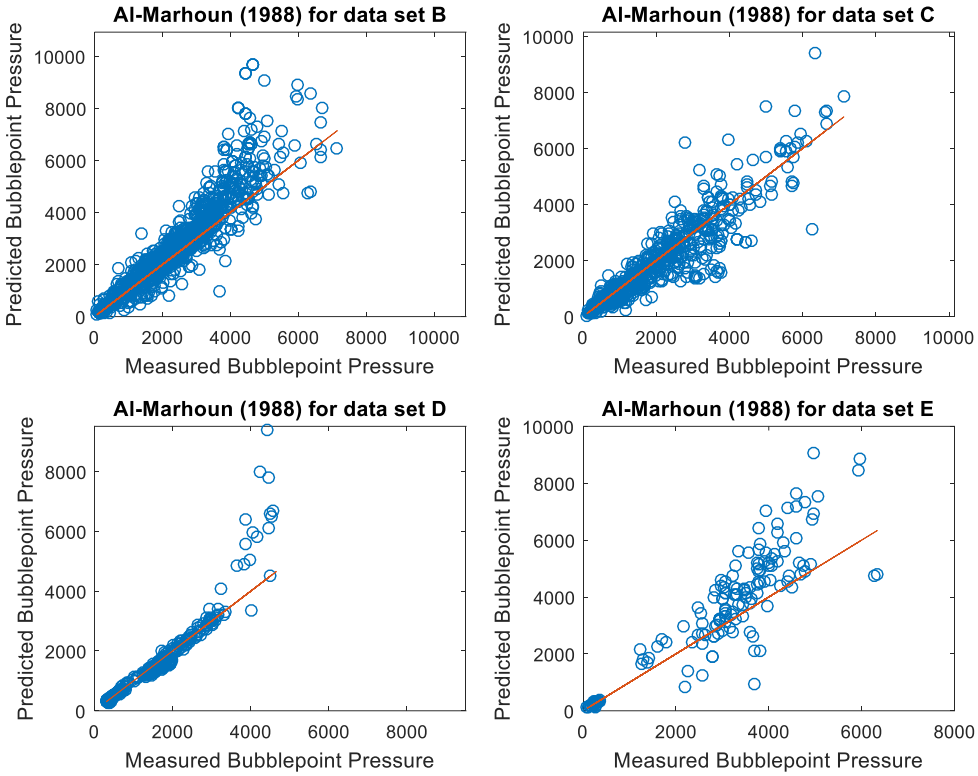


Figure 5.45. Cross plot for bubblepoint pressure with Al-Marhoun (1988)

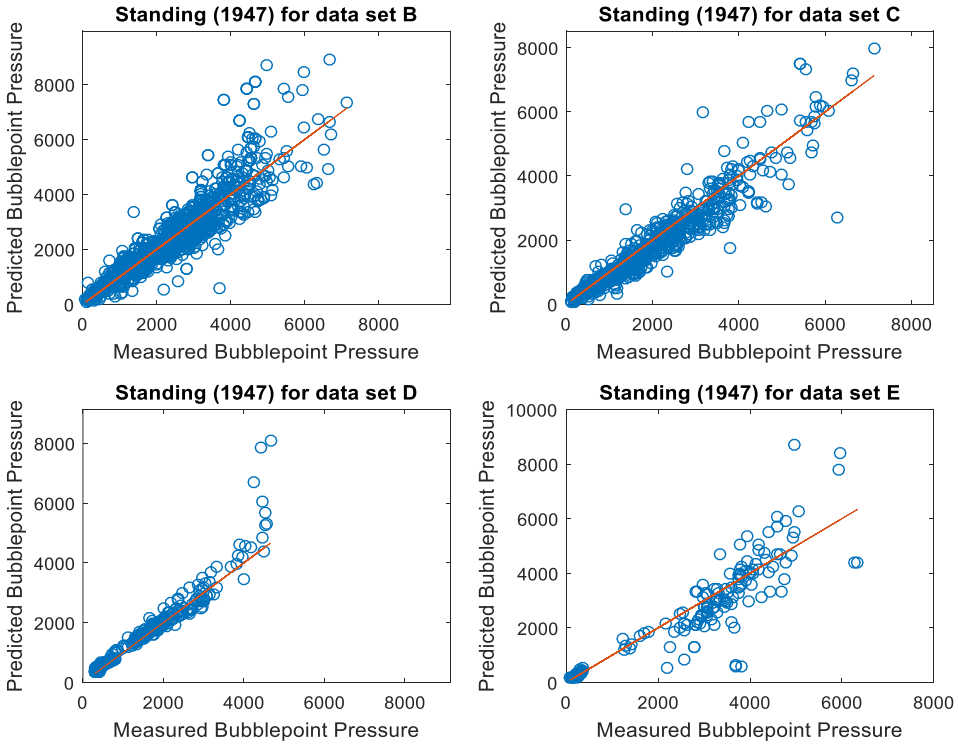


Figure 5.46. Cross plot for bubblepoint pressure with Standing (1947)

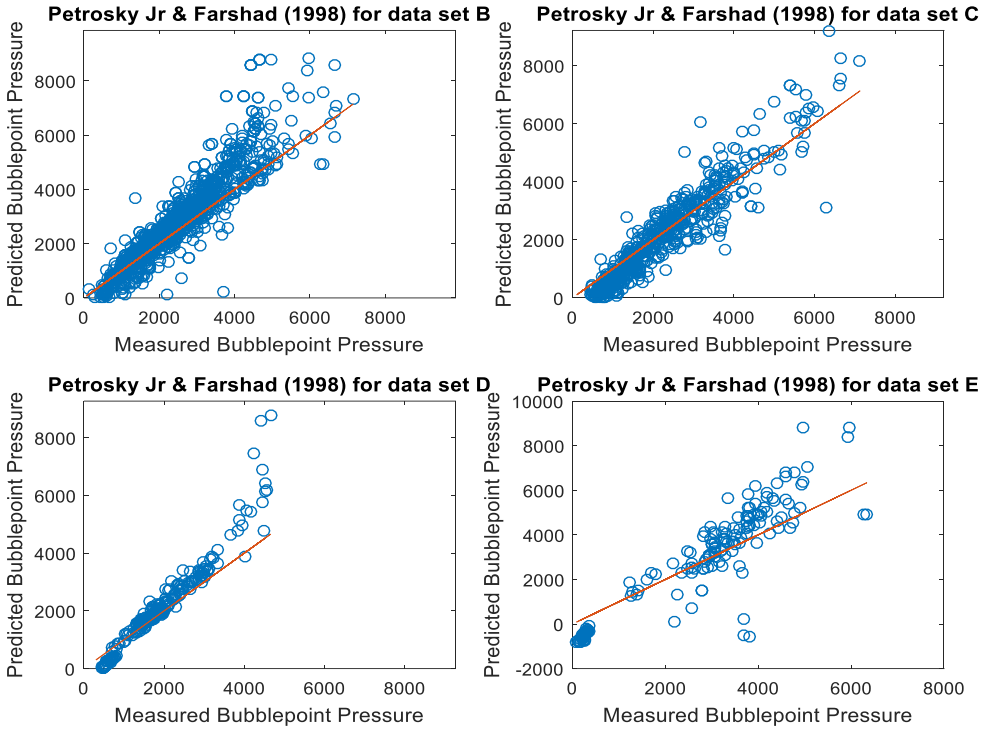


Figure 5.47. Cross plot for bubblepoint pressure with Petrosky Jr & Farshad (1998)

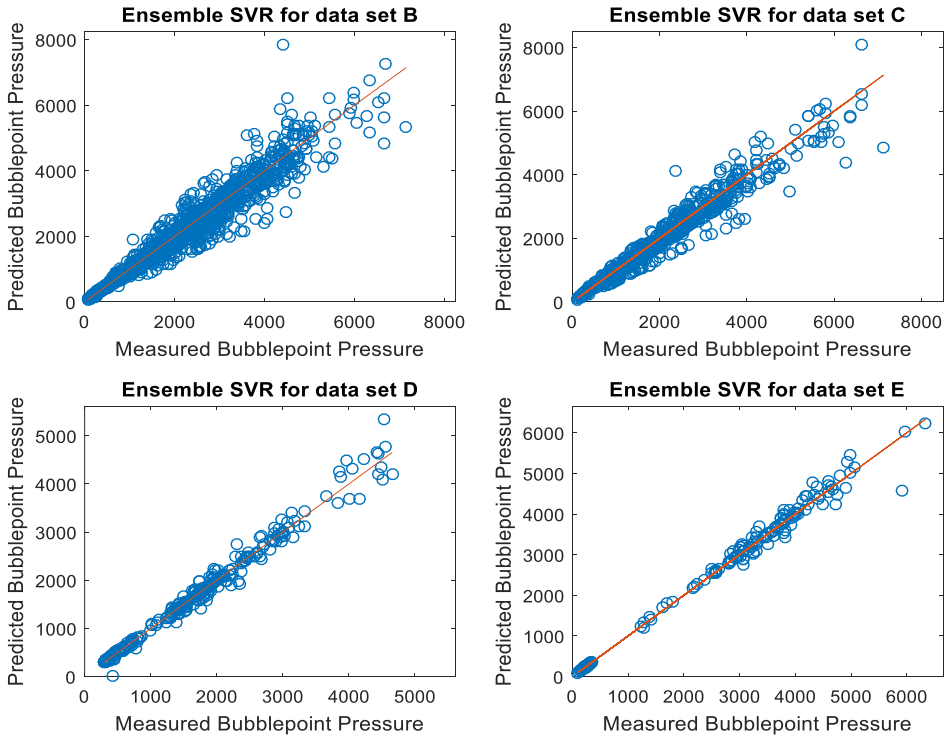


Figure 5.48. Cross plot for bubblepoint pressure with Ensemble SVR

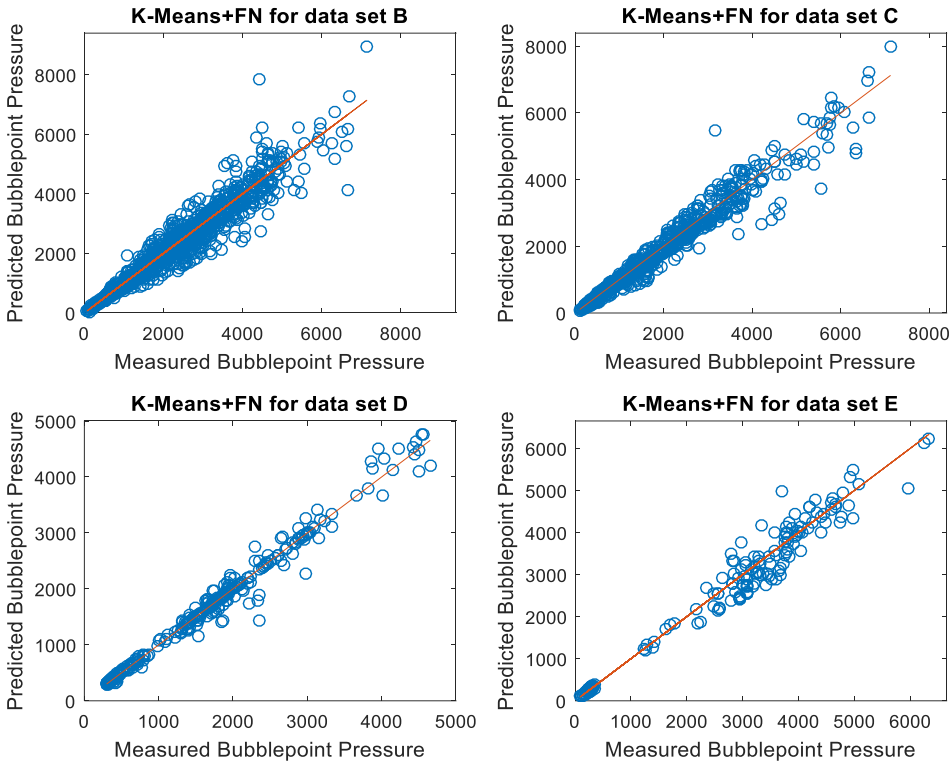


Figure 5.49. Cross plot for bubblepoint pressure with K-Means+FN

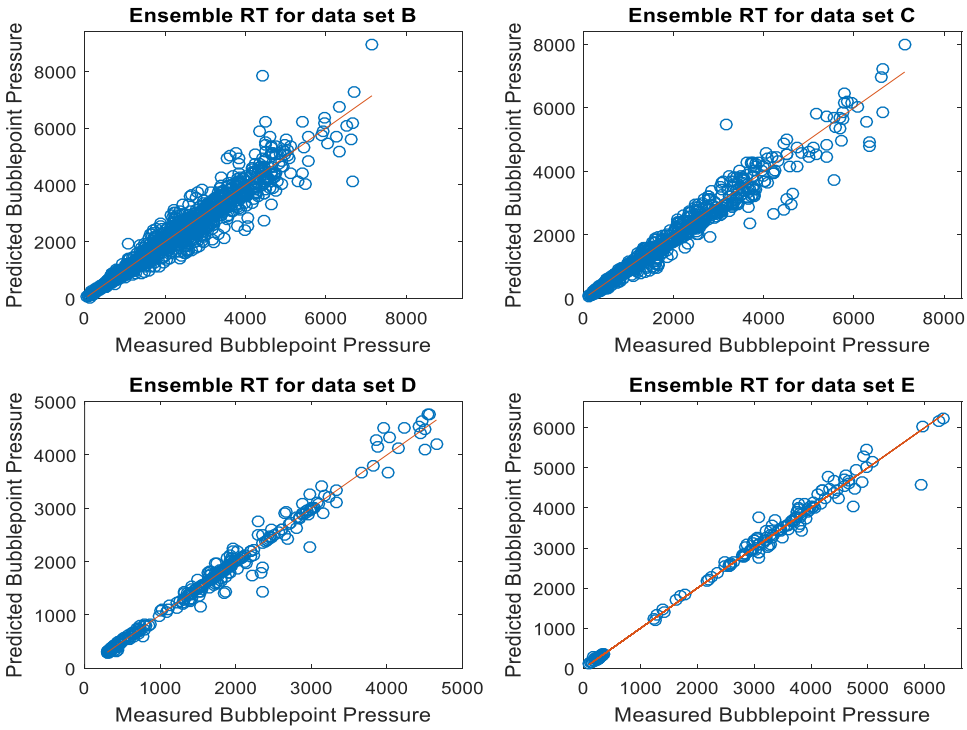


Figure 5.50. Cross plot for bubblepoint pressure with Ensemble RT

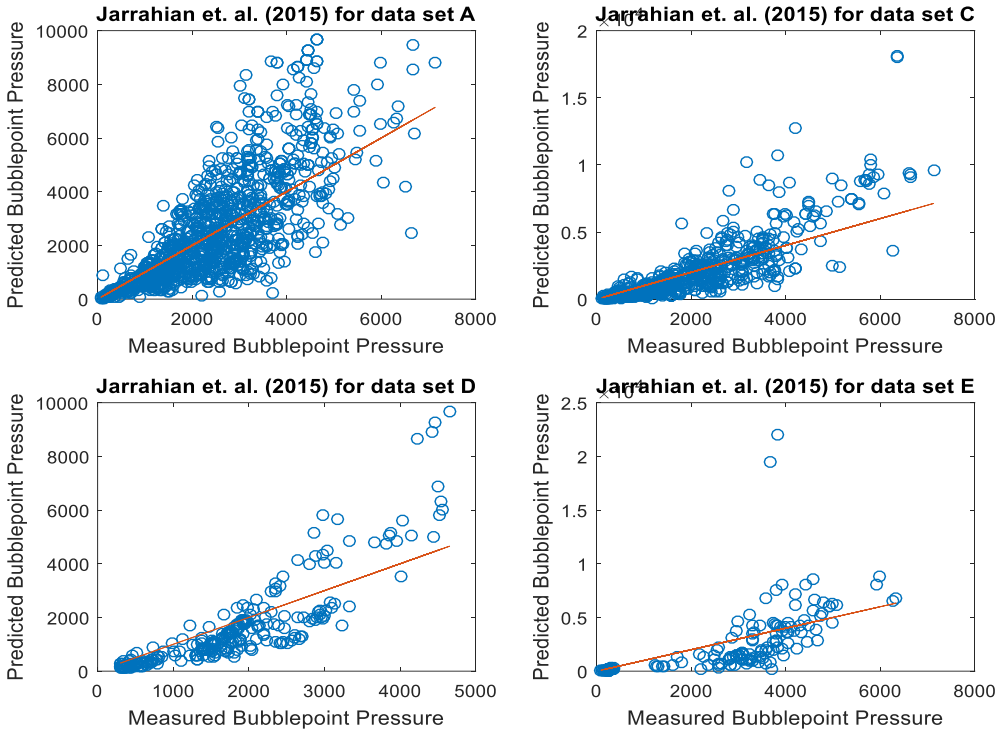


Figure 5.51. Cross plot for bubblepoint pressure with Jarrahian et al. (2015)

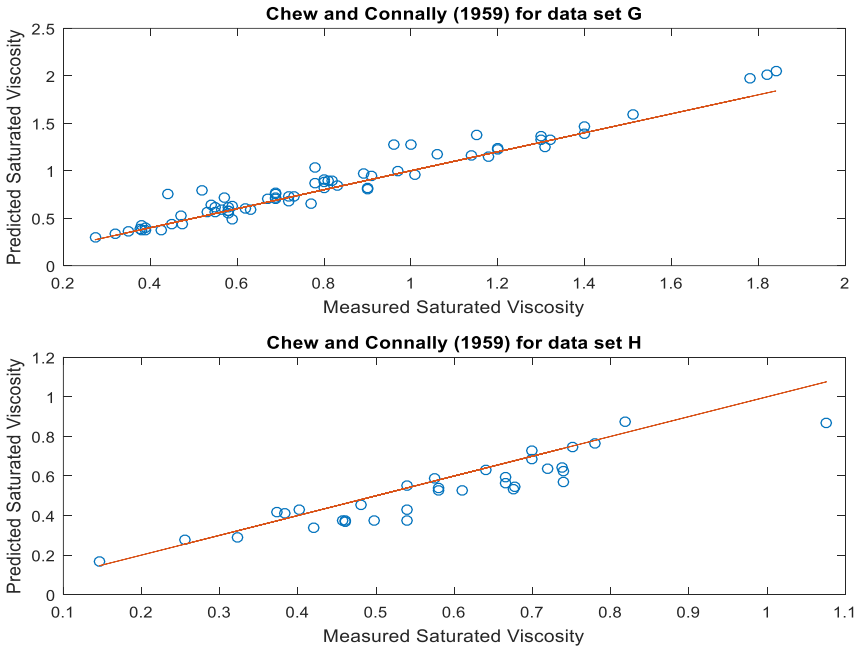


Figure 5.52. Cross plot for saturated viscosity with Chew & Connally (1959)

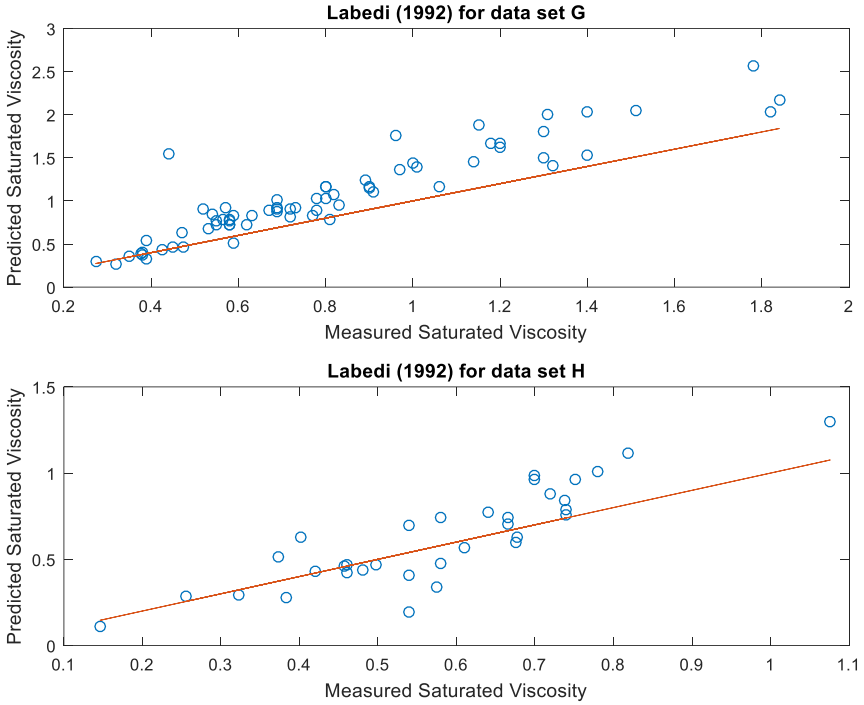


Figure 5.53. Cross plot for saturated viscosity with Labedi (1992)

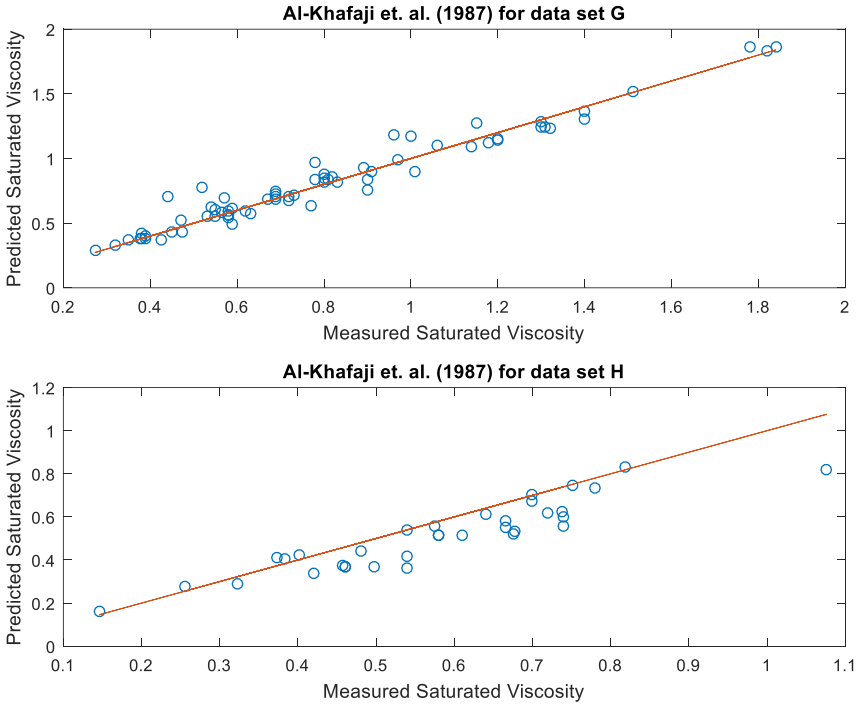


Figure 5.54. Cross plot for saturated viscosity with Al-Khafaji et. al. (1987)

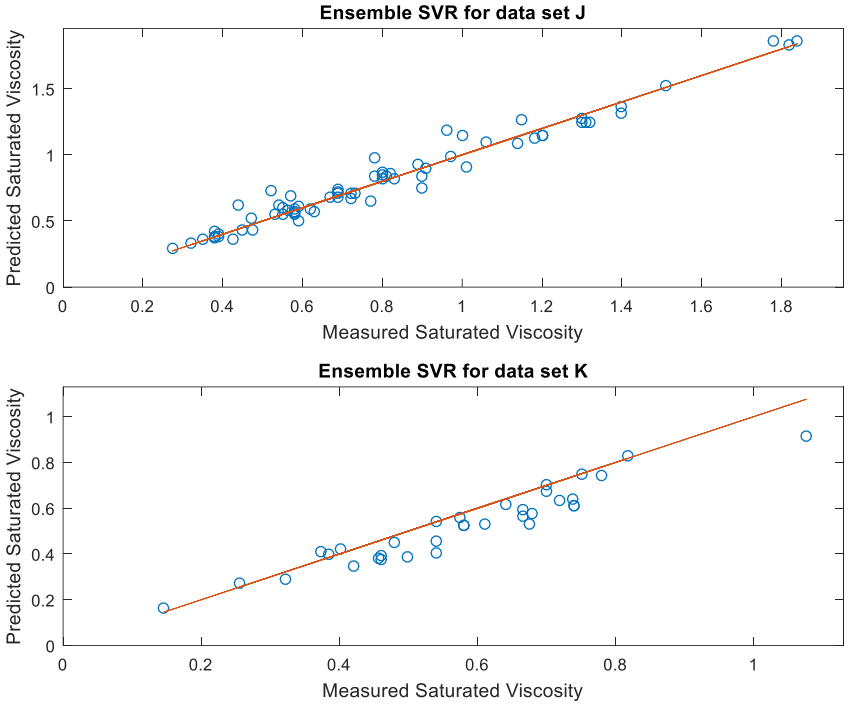


Figure 5.55. Cross plot for saturated viscosity with Ensemble SVR

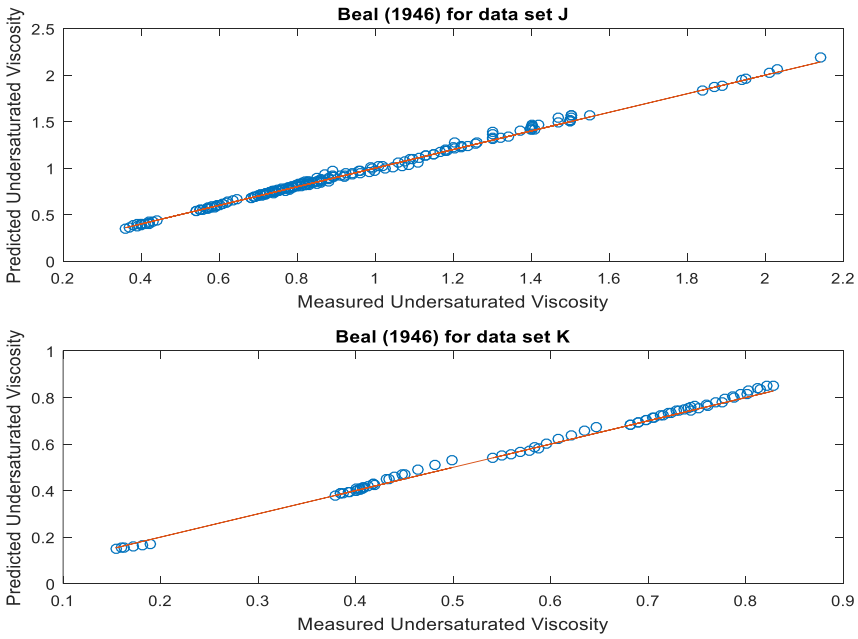


Figure 5.56. Cross plot for undersaturated viscosity with Beal (1946)

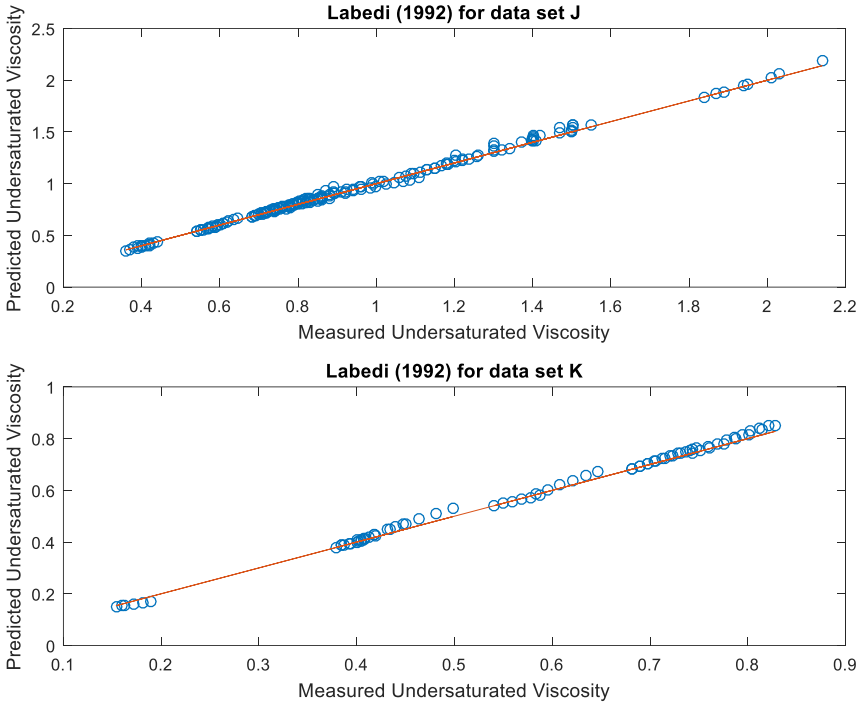


Figure 5.57. Cross plot for undersaturated viscosity with Labedi (1992)

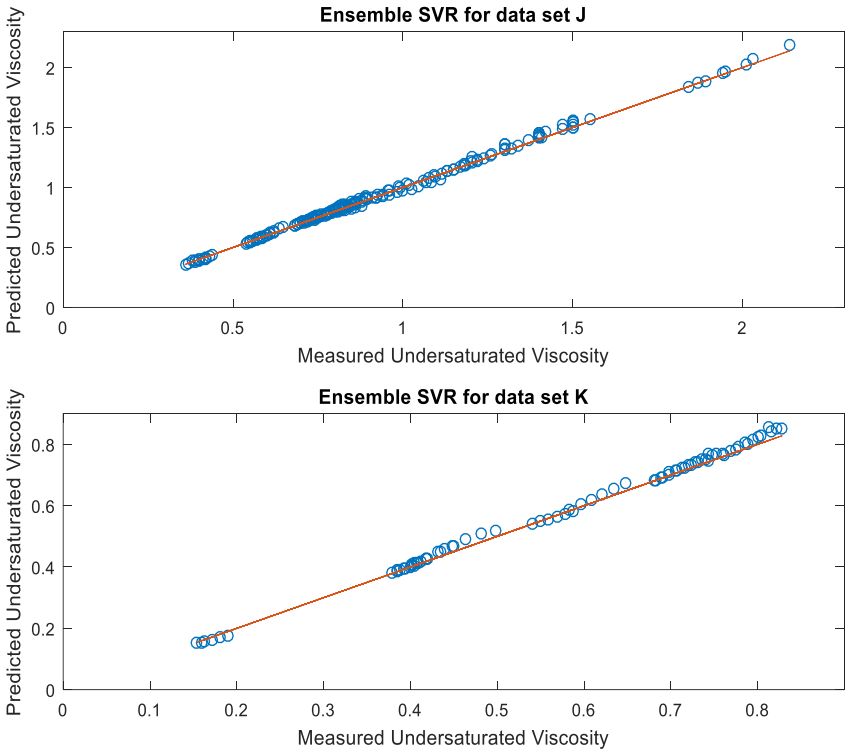


Figure 5.58. Cross plot for undersaturated viscosity with Ensemble SVR

Chapter 6. Conclusion and Future Work

6.1. Concluding Remarks

There is always a quest for improving the prediction of reservoir fluid properties. This is mainly due to their relevance, importance and applications in the petroleum industry. In view of this, relevant existing empirical correlations and machine learning or computational intelligence techniques that have been used for reservoir fluid characterisation are reviewed. However, only standalone models of these computational intelligence techniques or their hybrid systems have been used. Also, some research works have used API grouping to improve the performance of their empirical correlations. In light of these, ensemble systems and intelligent API clustering have been explored in this study to examine their possibility of improving the prediction of reservoir fluid properties.

6.2. Research Contributions

A new hybrid system inspired by API gravity grouping of petroleum fluids has been developed. The hybrid system uses K-means clustering for natural grouping of the input data set before performing the prediction by functional network. The hybrid functional network in most cases performs better than the standalone functional network though the margin could be sometimes low.

An innate capability of the functional network, not obviously displayed by other explored machine learning (ML) techniques, is the ability to suppress non-correlating variables in the input data set. This makes the functional network to serve both as a feature selection method as well as prediction technique. This characteristic was revealed and evident in the prediction of dead oil viscosity.

Also, two different ensemble systems using support vector regression (SVR) and regression trees (RTs) have been developed. The selection of the base models for the ensemble systems has been done using a method tagged ‘tying ranking’.

Two different experimentations have been performed: the first looks into the impact of training the machine learning algorithms with diverse data sets of different sizes and the second experimentation observes the effect of diverse correlating variables for the same target prediction output. The former was carried out using bubblepoint pressure and saturated oil formation volume factor while the latter was on dead oil, saturated and undersaturated viscosity.

Chapter 4 shows the steps for implementing the three algorithms. The “Tying ranking” technique which has been introduced to model the ensemble support vector regression and ensemble regression tree is explained in chapter.

In chapter 5, details of the results from the experimental work have been highlighted and discussed. Mainly, four different statistical measures (correlation coefficient, root mean squared error, average absolute percent relative error and maximum absolute percent relative error) have been used to compare and evaluate the developed three ML models, their standalone techniques and the relevant empirical correlations.

Summary reports of the minimum and maximum values of these statistical measures and their average values across different testing data sets have also been reported in chapter 5. At least of the developed newly developed ML techniques has best overall performance for each of the examined reservoir fluid properties. Also, most ML techniques have more generalisation capabilities than the empirical correlations especially when the summary reports of the statistical measures for all the examined reservoir fluid properties are analysed.

Different factors can influence the performances of an ML model. Diversification and quantity of data set, model input feature selection and determination of optimal learning parameters are notable factors that influence the performances of an ML model.

Training an ML technique with a large data sets most likely help in arriving at a more generic and consistent model. This has been examined in the experimental work involving the prediction of bubblepoint pressure and saturated oil formation volume factor. Also, careful determination of the correlating input variables will also go a long way in arriving at a robust ML technique, rather than introducing polluting inputs. Likewise, intelligent clustering or grouping of input data sets may improve the prediction of given PVT property and this will likely work better with large input data sets.

Lastly, the impacts of different statistical measures used to rate and evaluate methods in this study have been examined. The erroneous usage of standard deviation by some previous research works have been unveiled. Sample cross plots have also been shown to relate the influence and impacts of these statistical measures.

6.2. Future Work

The opportunities in ML research and applications are quite immense and not exhaustive. The following are some considerations that could possibly further the research in this work.

- One of the notable problems in the results of the ensemble SVR is that E_{max} is sometimes comparatively very high even when its other evaluation metric measures are competitive or even outperform other ML techniques. This could be a result of outliers in some

datasets. Possibility of using outlier detection techniques to pre-process the data before training is desired. This may also improve the results of other ML techniques.

- Also, some of the data used in this work are from pre 1990 PVT reports. However, some PVT experts have indicated that better quality data can be acquired from post 1990 PVT reports as they have benefited from technology advancement in data acquisition and PVT laboratory analysis. While access to many post 1990 PVT data has been difficult due to the current economy turmoil in the petroleum industry, possibly orchestrated by the current protracted relatively low crude oil price, using recent and modernly acquired PVT data may improve the prediction results appreciably.
- In the same vein, the potential for intelligent grouping or clustering introduced in this work can be explored further. With availability of larger data sets, other clustering techniques can be explored. Also, more machine learning techniques can be hybridized with different clustering techniques.
- In the “Tying Ranking” that was used to select the ensemble base models, a weighting function with respect to the rank of the base models can be used to determine their contribution to the final prediction rather than using the average.

References

Al-Khafaji, A. H., Abdul-Majeed, G. H., & Hassoon, S. F. (1987). Viscosity correlation for dead, live and undersaturated crude oils. *J. Pet. Res*, *6*(2), 1–16.

Al-Marhoun, M. A. (1988). PVT correlations for Middle East crude oils. *Journal of Petroleum Technology*, *40*(05), 650–666.

Al-Marhoun, M. A. (1992). New Correlation for formation Volume Factor of oil and gas Mixtures. *Journal of Canadian Petroleum Technology*, *31*(3), 22–26.

Al-Marhoun, M. A. (2003). The coefficient of isothermal compressibility of black oils. In *Middle East Oil Show*. Society of Petroleum Engineers.

Al-Marhoun, M. A. (2004). Evaluation of empirically derived PVT properties for Middle East crude oils. *Journal of Petroleum Science and Engineering*, *42*(2), 209–221.

Al-Marhoun, M. A. (2009). The Oil Compressibility below Bubble Point Pressure Revisited-Formulations and Estimations. In *SPE Middle East Oil and Gas Show and Conference*. Society of Petroleum Engineers.

Al-Marhoun, M. A., & Osman, E. A. (2002). Using artificial neural networks to develop new PVT correlations for Saudi crude oils. In *Abu Dhabi international petroleum exhibition and conference*. Society of Petroleum Engineers.

Al-Shammasi, A. A. (2001). A review of bubblepoint pressure and oil formation volume factor correlations. *SPE Reservoir Evaluation & Engineering*, *4*(02), 146–160.

Ali, J. K. (1994). Neural networks: a new tool for the petroleum industry? In *European petroleum computer conference*. Society of Petroleum Engineers.

Almehaideb, R. A. (1997). Improved PVT correlations for UAE crude oils. In *Middle east oil show and conference*. Society of Petroleum Engineers.

Alomair, O., Elsharkawy, A., & Alkandari, H. (2014). A viscosity prediction model for Kuwaiti heavy crude oils at elevated temperatures. *Journal of Petroleum Science and Engineering*, *120*, 102–110.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Arabloo, M., Amooie, M.-A., Hemmati-Sarapardeh, A., Ghazanfari, M.-H., & Mohammadi, A. H. (2014). Application of constrained multi-variable search methods for prediction of PVT properties of crude oil systems. *Fluid Phase Equilibria*, *363*, 121–130.

Asafa, T. B., Adeniran, A. A., & Olatunji, S. O. (2015). Functional networks models for rapid characterization of thin films: An application to ultrathin polycrystalline silicon germanium films. *Applied Soft Computing*, *28*, 11–18.

Asoodeh, M., & Bagheripour, P. (2012). Estimation of bubble point pressure from PVT data using a power-law committee with intelligent systems. *Journal of Petroleum Science and Engineering*, *90*, 1–11.

Ayoub, M. A., Raja, D. M., & Al-Marhoun, M. A. (2007). Evaluation of below bubble point viscosity correlations & construction of a new neural network model.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31.

Beal, C. (1946). The viscosity of air, water, natural gas, crude oil and its associated gases at oil field temperatures and pressures. *Transactions of the AIME*, *165*(01), 94–115.

Beggs, H. D., & Robinson, J. R. (1975). Estimating the viscosity of crude oil systems. *Journal of Petroleum Technology*, *27*(09), 1140–1149.

Bello, O. O., Reinicke, K. M., & Patil, P. A. (2008). Comparison of the performance of empirical models used for the prediction of the PVT properties of crude oils of the niger delta. *Petroleum Science and Technology*, *26*(5), 593–609.

Bennison, T. (1998). Prediction of heavy oil viscosity. In *Presented at the IBC Heavy Oil Field Development Conference* (Vol. 2, p. 4).

Bergman, D. F., & Sutton, R. P. (2007). An update to viscosity correlations for gas-saturated crude oils. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Bergman, D. F., & Sutton, R. P. (2009). A consistent and accurate dead-oil-viscosity method. *SPE Reservoir Evaluation & Engineering*, *12*(06), 815–840.

Boukadi, F. H., Bemani, A. S., & Hashmi, A. (2002). PVT empirical models for saturated Omani crude oils. *Petroleum Science and Technology*, *20*(1–2), 89–100.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Calhoun Jr, J. C. (1947). *Fundamentals of reservoir engineering*. Norman, Oklahoma, University of Oklahoma Press, 35.

Carbonell, J., Michalski, R., & Mitchell, T. (1983). An Overview of Machine Learning. *Machine Learning SE - 1*. http://doi.org/10.1007/978-3-662-12405-5_1

Castillo, E., Cobo, A., Gutiérrez, J. M., & Pruneda, E. (2000). Functional Networks: A New Network-Based Methodology. *Computer-Aided Civil and Infrastructure Engineering*, *15*(2), 90–106.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y., & Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, 143–167.

Chew, J., & Connally, C. A. (1959). A viscosity correlation for gas-saturated crude oils. *Trans. AIME*, *216*, 23–25.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dandekar, A. Y. (2013). *Petroleum reservoir rock and fluid properties*. CRC press.

De Ghetto, G., Paone, F., & Villa, M. (1995). Pressure-volume-temperature correlations for heavy and extra heavy oils. In *SPE International Heavy Oil Symposium*. Society of Petroleum Engineers.

De Ghetto, G., & Villa, M. (1994). Reliability analysis on PVT correlations. In *European Petroleum Conference*. Society of Petroleum Engineers.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple classifier systems* (pp. 1–15). Springer.

Dindoruk, B., & Christman, P. G. (2004). PVT Properties and Viscosity Correlations for Gulf of Mexico Oils. *SPE Reservoir Evaluation & Engineering*, *7*(06), 427–437.

Dokla, M., & Osman, M. (1992). Correlation of PVT Properties for UAE Crudes (includes associated papers 26135 and 26316). *SPE Formation Evaluation*, *7*(01), 41–46.

Dusseldorp, E., & Meulman, J. J. (2004). The regression trunk approach to discover treatment covariate interaction. *Psychometrika*, *69*(3), 355–374.

Egbogah, E. O., & Ng, J. T. (1990). An improved temperature-viscosity correlation for crude oil systems. *Journal of Petroleum Science and Engineering*, *4*(3), 197–200.

El-hoshoudy, A. N., Farag, A. B., Ali, O. I. M., El-Batanoney, M. H., Desouky, S. E. M., & Ramzi, M. (2013). New correlations for prediction of viscosity and density of Egyptian oil reservoirs. *Fuel*, *112*, 277–282.

El-Sebakhy, E. A., Sheltami, T., Al-Bokhitan, S. Y., Shaaban, Y., Raharja, P. D., & Khaeruzzaman, Y. (2007). Support vector machines framework for predicting the PVT properties of crude oil systems. In *SPE Middle East Oil and Gas Show and Conference*. Society of Petroleum Engineers.

Elsebakhi, E., Lee, F., Schendel, E., Haque, A., Kathireason, N., Pathare, T., … Al-Ali, R. (2015). Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. *Journal of Computational Science*, *11*, 69–81.

Elsharkawy, A. M. (1998). Modeling the properties of crude oil and gas systems using RBF network. In *SPE Asia Pacific Oil and Gas Conference and Exhibition*. Society of Petroleum Engineers.

Elsharkawy, A. M. (2003). An empirical model for estimating the saturation pressures of crude oils. *Journal of Petroleum Science and Engineering*, *38*(1), 57–77.

Elsharkawy, A. M., & Alikhan, A. A. (1997). Correlations for predicting solution gas/oil ratio, oil formation volume factor, and undersaturated oil compressibility. *Journal of Petroleum Science and Engineering*, *17*(3), 291–302.

Elsharkawy, A. M., & Alikhan, A. A. (1999). Models for predicting the viscosity of Middle East crude oils. *Fuel*, *78*(8), 891–903.

Elsharkawy, A. M., Hassan, S. A., Hashim, Y. S. K., & Fahim, M. A. (2003). New compositional models for calculating the viscosity of crude oils. *Industrial & Engineering Chemistry Research*, *42*(17),

4132–4142.

Farshad, F. F., Garber, J. D., & Lorde, J. N. (2000). Predicting temperature profiles in producing oil wells using artificial neural networks. *Engineering Computations*, *17*(6), 735–754.

Farshad, F., LeBlanc, J. L., Garber, J. D., & Osorio, J. G. (1996). Empirical PVT correlations for Colombian crude oils. In *SPE Latin America/Caribbean petroleum engineering conference*. Society of Petroleum Engineers.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *ICML* (Vol. 96, pp. 148–156).

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Gharbi, R. B., & Elsharkawy, A. M. (1997). Neural network model for estimating the PVT properties of Middle East crude oils. In *Middle East Oil Show and Conference*. Society of Petroleum Engineers.

Gharbi, R. B., Elsharkawy, A. M., & Karkoub, M. (1999). Universal neural-network-based model for estimating the PVT properties of crude oil systems. *Energy & Fuels*, *13*(2), 454–458.

Ghorbani, B., Hamedi, M., Shirmohammadi, R., Mehrpooya, M., & Hamedi, M.-H. (2016). A novel multi-hybrid model for estimating optimal viscosity correlations of Iranian crude oil. *Journal of Petroleum Science and Engineering*, *142*, 68–76.

Glasø, Ø. (1980). *Generalized Pressure-Volume-Temperature Correlations. JPT 32 (5) 785–795*. SPE-8016-PA. DOI: 10.2118/8016-PA.

Goda, H. M., Shokir, E.-M., Eissa, M., Fattah, K. A., & Sayyouh, M. H. (2003). Prediction of the PVT data using neural network computing theory. In *Nigeria Annual International Conference and Exhibition*. Society of Petroleum Engineers.

Hajizadeh, Y. (2007). Viscosity prediction of crude oils with genetic algorithms. In *Latin American & Caribbean Petroleum Engineering Conference*. Society of Petroleum Engineers.

Hanafy, H. H., Macary, S. M., ElNady, Y. M., Bayomi, A. A., & El Batanony, M. H. (1997). A new approach for predicting the crude oil properties. In *SPE Production Operations Symposium*. Society of Petroleum Engineers.

Hemmati, M. N., & Kharrat, R. (2007). A correlation approach for prediction of crude oil PVT properties. In *SPE Middle East Oil and Gas Show and Conference*. Society of Petroleum Engineers.

Hossain, M. S., Sarica, C., Zhang, H.-Q., Rhyne, L., & Greenhill, K. L. (2005). Assessment and development of heavy oil viscosity correlations. In *SPE International Thermal Operations and Heavy Oil Symposium*. Society of Petroleum Engineers.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Ikiensikimama, S. S., & Ajienka, J. A. (2012). Impact of PVT correlations development on

hydrocarbon accounting: the case of the Niger Delta. *Journal of Petroleum Science and Engineering*, *81*, 80–85.

Jang, J.-S. R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *Systems, Man and Cybernetics, IEEE Transactions On*, *23*(3), 665–685.

Jarrahian, A., Moghadasi, J., & Heidaryan, E. (2015). Empirical estimating of black oils bubblepoint (saturation) pressure. *Journal of Petroleum Science and Engineering*, *126*, 69–77.

Kartoatmodjo, T. R. S., & Schmidt, Z. (1991). New correlations for crude oil physical properties.

Kartoatmodjo, T., & Schmidt, Z. (1994). Large data bank improves crude physical property correlations. *Oil and Gas Journal;(United States)*, *92*(27).

Kecman, V. (2001). *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. Bradford Book. Retrieved from https://books.google.co.uk/books?id=W5SAhUqBVYoC

Khairy, M., El-Tayeb, S., & Hamdallah, M. (1998). PVT correlations developed for Egyptian crudes. *Oil and Gas Journal*, *96*(18).

Khamehchi, E., Rashidi, F., Rasouli, H., & Ebrahimian, A. (2009). Novel empirical correlations for estimation of bubble point pressure, saturated viscosity and gas solubility of crude oils. *Petroleum Science*, *6*(1), 86–90.

Khan, S. A., Al-Marhoun, M. A., Duffuaa, S. O., & Abu-Khamsin, S. A. (1987). Viscosity correlations for Saudi Arabian crude oils. In *Middle East Oil Show*. Society of Petroleum Engineers.

Khoukhi, A. (2012). Hybrid soft computing systems for reservoir PVT properties prediction. *Computers & Geosciences*, *44*, 109–119.

Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*(454), 589–604.

Kim, H., & Loh, W.-Y. (2003). Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics*, *12*(3), 512–530.

Labedi, R. (1992). Improved correlations for predicting the viscosity of light crudes. *Journal of Petroleum Science and Engineering*, *8*(3), 221–234.

Lasater, J. A. (1958). Bubble point pressure correlation. *Journal of Petroleum Technology*, *10*(05), 65–67.

Lior Rokach, O. M. (2015). *Data Mining With Decision Trees: Theory and Applications (2nd Edition)*. *Series in Machine Perception and Artifical Intelligence*. World Scientific Publishing Company.

Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, 361–386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, 1710–1737.

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*,

815–840.

Loh, W.-Y., & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, *83*(403), 715–725.

Loh, W. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Loh, W. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348.

Mahmood, M. A., & Al-Marhoun, M. A. (1996). Evaluation of empirically derived PVT properties for Pakistani crude oils. *Journal of Petroleum Science and Engineering*, *16*(4), 275–290.

Malallah, A., Gharbi, R., & Algharaib, M. (2006). Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation. *Energy & Fuels*, *20*(2), 688–698.

McCain Jr, W. D., Rollins, J. B., & Lanzi, A. J. V. (1988). The coefficient of isothermal compressibility of black oils at pressures below the bubblepoint. *SPE Formation Evaluation*, *3*(03), 659–662.

McCain, W. D. (1990). *The properties of petroleum fluids*. PennWell Books.

Messenger, R., & Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, *67*(340), 768–772.

Mohaghegh, S. (2000). Virtual-intelligence applications in petroleum engineering: Part 1—Artificial neural networks. *Journal of Petroleum Technology*, *52*(09), 64–73.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*(302), 415–434.

Naseri, A., Nikazar, M., & Dehghani, S. A. M. (2005). A correlation approach for prediction of crude oil viscosities. *Journal of Petroleum Science and Engineering*, *47*(3), 163–174.

Ng, J. T. H., & Egbogah, E. O. (1983). An improved temperature-viscosity correlation for crude oil systems. In *Annual Technical Meeting*. Petroleum Society of Canada.

Nguyen, H. H., & Chan, C. W. (2005). Applications of data analysis techniques for oil production prediction. *Engineering Applications of Artificial Intelligence*, *18*(5), 549–558.

Obomanu, D. A., & Okpobiri, G. A. (1987). Correlating the PVT properties of Nigerian crudes. *Journal of Energy Resources Technology*, *109*(4), 214–217.

Olatunji, S. O., Selamat, A., & Raheem, A. A. A. (2010). Modeling permeability prediction using extreme learning machines. In *Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010 Fourth Asia International Conference on* (pp. 29–33). IEEE.

Olatunji, S. O., Selamat, A., Raheem, A. A. A., & Omatu, S. (2011). Modeling the correlations of crude oil properties based on sensitivity based linear learning method. *Engineering Applications of Artificial Intelligence*, *24*(4), 686–696.

Oloso, M. A., Hassan, M. G., Bader-El-Den, M. B., & Buick, J. M. (2017). Hybrid functional networks

for oil reservoir PVT characterisation. *Expert Systems with Applications*, *87*. http://doi.org/10.1016/j.eswa.2017.06.014

Oloso, M. A., Hassan, M. G., Bader-El-Den, M. B., & Buick, J. M. (2018). Ensemble SVM for characterisation of crude oil viscosity. *Journal of Petroleum Exploration and Production Technology*, *8*(2). http://doi.org/10.1007/s13202-017-0355-x

Oloso, M. A., Khoukhi, A., Abdulraheem, A., & Elshafei, M. (2009). Prediction of crude oil viscosity and gas/oil ratio curves using recent advances to neural networks. In *Society of Petroleum Engineers - SPE/EAGE Reservoir Characterization and Simulation Conference ,19-21 October, Abu Dhabi, UAE* (Vol. 1).

Omar, M. I., & Todd, A. C. (1993). Development of new modified black oil correlations for Malaysian crudes. In *SPE Asia Pacific oil and gas conference*. Society of Petroleum Engineers.

Osman, E.-S. A., & Al-Marhoun, M. A. (2005). Artificial neural networks models for predicting PVT properties of oil field brines. In *SPE Middle East Oil and Gas Show and Conference*. Society of Petroleum Engineers.

Osman, E. A., & Abdel-Aal, R. E. (2002). Abductive networks: a new modeling tool for the oil and gas industry. *Paper SPE*, *77882*, 8–10.

Osman, E. A., Abdel-Wahhab, O. A., & Al-Marhoun, M. A. (2001). Prediction of oil PVT properties using neural networks. In *SPE Middle East Oil Show*. Society of Petroleum Engineers.

Ostermann, R. D., & Owolabi, O. O. (1983). Correlations for the reservoir fluid properties of Alaskan crudes. In *SPE California Regional Meeting*. Society of Petroleum Engineers.

Peng, D.-Y., & Robinson, D. B. (1976). A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals*, *15*(1), 59–64.

Petrosky Jr, G. E., & Farshad, F. (1998). Pressure-volume-temperature correlations for Gulf of Mexico crude oils. *SPE Reservoir Evaluation & Engineering*, *1*(05), 416–420.

Petrosky Jr, G. E., & Farshad, F. F. (1995). Viscosity correlations for Gulf of Mexico crude oils. In *SPE Production Operations Symposium*. Society of Petroleum Engineers.

Quinlan, J. R. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343–348). Singapore.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Redlich, O., & Kwong, J. N. S. (1949). On the thermodynamics of solutions. V. An equation of state. Fugacities of gaseous solutions. *Chemical Reviews*, *44*(1), 233–244.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

Soave, G. (1972). Equilibrium constants from a modified Redlich-Kwong equation of state. *Chemical Engineering Science*, *27*(6), 1197–1203.

Standing, M. B. (1947). A pressure-volume-temperature correlation for mixtures of California oils and gases. In *Drilling and Production Practice*. American Petroleum Institute.

Standing, M. B. (1962). Oil-system correlations. *Petroleum Production Handbook*, *2*.

Sutton, R. P. (2008). An Accurate Method for Determining Oil PVT Properties Using the Standing-Katz Gas Z-Factor Chart. *SPE Reservoir Evaluation & Engineering*, *11*(02), 246–266.

Talebi, R., Ghiasi, M. M., Talebi, H., Mohammadyian, M., Zendehboudi, S., Arabloo, M., & Bahadori, A. (2014). Application of soft computing approaches for modeling saturation pressure of reservoir oils. *Journal of Natural Gas Science and Engineering*, *20*, 8–15.

Twu, C. H. (1985). Internally consistent correlation for predicting liquid viscosities of petroleum fractions. *Industrial & Engineering Chemistry Process Design and Development*, *24*(4), 1287–1293.

Valko, P. P., & McCain, W. D. (2003). Reservoir oil bubblepoint pressures revisited; solution gas–oil ratios and surface gas specific gravities. *Journal of Petroleum Science and Engineering*, *37*(3), 153–169.

Van-der-Waals. (1873). *On the continuity of gas and liquid state*. *Doctoral Dissertation, Leiden University*.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=EqgACAAAQBAJ&oi=fnd&pg=PR7&ots=g3DZmBbV11&sig=C-sIsGhXGgpv_nwjFOAd5SBnzaY

Varotsis, N., Gaganis, V., Nighswander, J., & Guieze, P. (1999). A novel non-iterative method for the prediction of the PVT behavior of reservoir fluids. In *SPE annual technical conference*.

Vazquez, M., & Beggs, H. D. (1980). Correlations for fluid physical property prediction. *Journal of Petroleum Technology*, *32*(06), 968–970.

Velarde, J., Blasingame, T. A., & McCain Jr, W. D. (1997). Correlation of black oil properties at pressures below bubble point pressure-A new approach. In *Annual Technical Meeting*. Petroleum Society of Canada.

Velarde, J., Blosingame, T. A., & McCain Jr, W. D. (1999). Correlation of Black Oil Properties at Pressures Below Bubble Point Pressure--A New Approach (97-93). *Journal of Canadian Petroleum Technology*, *38*(13), 62.

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

Appendix I. Statistical Measures and Statistical Descriptions of the Data Sets

A. Mathematical equations of the statistical measures for the performance Analysis in this work are presented below.

A.1. Average percent relative error

$$E_r = \frac{1}{n}\sum_1^n E_i \quad \text{(A-1)}$$

Where,

$$E_i = \left(\frac{X_{exp} - X_{pred}}{X_{exp}}\right)_i \times 100 \quad \text{(A-2)}$$

$$i = 1,2, \ldots . , n$$

A.2. Average absolute percent relative error

$$E_a = \frac{1}{n}\sum_1^n |E_i| \quad \text{(A-3)}$$

A.3. Maximum absolute percent relative error

$$E_{max} = \max_i |E_i| \quad \text{(A-4)}$$

A.4. Standard Deviation

$$SD = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^n (E_i - E_r)^2} \quad \text{(A-5)}$$

Where,

$E_r = \frac{1}{n}\sum_{i=1}^n E_i$.

A.5. Root mean squared

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^n E_i\right]^{0.5} \quad \text{(A-6)}$$

B. Statistical descriptions of data sets used for experimental work are presented in this section. The statistical attributes that have been presented are mean, minimum, maximum and standard deviations of the data points. The number of data points in each data set is also stated.

Table A.1. Statistical Description of Data Set A (895 data points)

Variable	Mean	Min	Max	SD
GOR	628.5659	10	2315.	398.0635
γ_g	0.8409	0.5560	1.367	0.1591
γ_{API}	36.4767	11.6	55	7.0214
T	176.4089	74	327	49.6316
P_b	2196.2053	70	6700	1170.3841
B_{ob}	1.3702	1.0301	2.4930	0.2259

Table A.2. Statistical Description of Data Set B (1200 data points)

Variable	Mean	Min	Max	SD
GOR	694.1222	10	2637	455.7028
γ_g	0.8198	0.556	1.367	0.1604
γ_{API}	36.1054	11.6	105.02	7.7312
T	179.6797	74	327	47.7760
P_b	2450.3906	70	7142.7	1276.0960
B_{ob}	1.4006	1	2.588	0.2553

Table A.3. Statistical Description of Data Set C (688 data points)

Variable	Mean	Min	Max	SD
GOR	532.4459	8.61	3617.27	503.3409
γ_g	0.8949	0.511	1.789	0.1859
γ_{API}	33.8755	11.4	63.7	8.7685
T	179.8747	58	341.6	53.8880
P_b	1993.9284	107.33	7127	1352.4236
B_{ob}	1.3300	1.028	2.887	0.2806

Table A.4. Statistical Description of Data Set D (273 data points)

Variable	Mean	Min	Max	SD
GOR	551.4286	56	2315	388.4553
γ_g	0.9588	0.61	1.384	0.1511
γ_{API}	32.9602	17.5	44.93	5.2030
T	151.1941	75	245	47.7782
P_b	1739.7132	300	4655	1017.9554
B_{ob}	1.3050	1.0432	2.49	0.1972

Table A.5. Statistical Description of Data Set E (170 data points)

Variable	Mean	Min	Max	SD
GOR	585.0233	24	2500	518.6863
γ_g	0.7886	0.57	1.8195	0.2444
γ_{API}	31.2019	15.9	115	11.6297
T	162.2878	85	247	35.4661
P_b	2654.3372	90	6336	1639.1786
B_{ob}	1.3110	1.0308	2.5	0.2794

Table A.6. Statistical Description of Data Set F (420 data points)

Variable	Mean	Min	Max	SD
P_b	1843.3838	381	7202	571.0408
γ_{API}	32.3747	23.5	63.7	3.1739
T	204.4242	130	240	24.2096
GOR	645.8687	184	1334	222.2733
γ_g	1.1301	0.8529	1.6313	0.1305
μ_{od}	7.0816	0.85	24.26	0.9611
μ_{ob}	1.7509	0.255	6.84	0.3338

Table A.7. Statistical Description of Data Set G (69 data points)

Variable	Mean	Min	Max	SD
P_b	1818.6618	371	10332	613.0885
γ_{API}	30.2985	20.9	60	3.7666
T	200.1765	120	235	25.1449
GOR	609.1765	180	1450	216.3569
γ_g	1.1240	0.8601	1.5212	0.1494
μ_{od}	4.3310	0.99	15.6	1.0302
μ_{ob}	1.0161	0.273	3.671	0.3707

Table A.8. Statistical Description of Data Set H (34 data points)

Variable	Mean	Min	Max	SD
P_b	2089.794	1167	4371	704.4843
γ_{API}	36.5	29.9	41.5	4.080107
T	216.5824	150	285	23.07177
GOR	797.2706	437	2405.7	341.5576
γ_g	1.136547	0.825	1.296	0.100951
μ_{od}	1.492147	0.6	2.923	0.45521
μ_{ob}	0.579529	0.146	1.076	0.182394

Table A.9. Statistical Description of Data Set I (1250 data points)

Variable	Mean	Min	Max	SD
P_b	1819.5885	381	3202	538.8896
T	206.5851	130	300	22.5964
P	2685.2587	500	4400	698.2372
μ_{od}	1.9878	0.85	5.99	0.9062
μ_{ob}	0.7327	0.255	1.84	0.3073
γ_{API}	28.8943	18.2	48	3.2576
μ_{oa}	0.7873	0.26	2.24	0.3374

Table A.10. Statistical Description of Data Set J (188 data points)

Variable	Mean	Min	Max	SD
P_b	1599.1702	741	2850	440.0104
T	200.8936	130	310	27.4325
P	2443.8670	867	4156	662.8027
μ_{od}	2.1192	0.85	4.09	0.8013
μ_{ob}	0.8684	0.35	1.84	0.3318
γ_{API}	30.3904	22.5	50	3.7531
μ_{oa}	0.9291	0.36	2.14	0.3591

Table A.11. Statistical Description of Data Set K (78 data points)

Variable	Mean	Min	Max	SD
P_b	2483.9872	1167	4371	1088.2991
T	216.4974	150	285	40.3038
P	3517.6667	1200	7515	1343.2258
μ_{od}	1.3696	0.6	2.923	0.5239
μ_{ob}	0.5353	0.146	0.751	0.1812
γ_{API}	37.3833	29.9	48	5.7401
μ_{oa}	0.5692	0.154	0.828	0.1887

Appendix II. Evaluated Empirical Correlations

A. Evaluated Correlation Coefficients for Bubblepoint Pressure

A.1. Standing (1947)

$$P_b = 18.2\left[\left(\frac{R_s}{\gamma_g}\right)^{0.83} 10^a - 1.4\right] \quad \text{(A-7)}$$

Where $a = 0.00091T - 0.0125\gamma_{API}$

A.2. Al-Marhoun (1988)

$$P_b = 5.38088 \times 10^{-3} R_s^{0.715082} \gamma_g^{-1.87784} \gamma_o^{3.1437} T^{1.32657} \quad \text{(A-8)}$$

A.3. Vazquez & Beggs (1980)

$$P_b = \left[A\left(\frac{R_s}{\gamma_{gc}}\right)10^a\right]^c \quad \text{(A-9)}$$

Where

$a = \frac{B\gamma_{API}}{T+459.67}$

$\gamma_{gc} = \gamma_g\left[1 + (0.5912 \times 10^{-4})\gamma_{API}\, T_{sp} log\left(\frac{P_{sp}}{114.7}\right)\right]$

P_{sp} and T_{sp} are separator pressure in psia and temperature in ^{0}F and the coefficients are in Table A

Table A.12. Coefficients for correlation of Vazquez & Beggs (1980)

Coefficient	$\gamma_{API} \leq 30$	$\gamma_{API} > 30$
A	27.64	56.06
B	-11.172	-10.393
C	0.9143	0.8425

A.4. Kartoatmodjo & Schmidt (1994)

For $API \leq 30$

$$P = ({}^{R_s}/_{A})^{0.9986} \quad \text{(A-10)}$$

Where

$A = 0.05958 \times \gamma_g^{0.7972} \times 10^a$

$$a = {}^{13.1405 \times \gamma_{API}}/_{(T+460)}$$

For $API > 30$

$$P = ({}^{R_s}/_{A})^{0.9143} \quad \text{(A-11)}$$

Where $A = 0.0315 \times \gamma_g^{0.7587} \times 10^a$

$$a = {}^{11.289 \times \gamma_{API}}/_{(T+460)}$$

A.5. Dokla & Osman (1992)

$$P_b = 0.836386 \times 10^4 \times \gamma_g^{-1.01049} \times \gamma_o^{0.10799} \times T^{-0.952584} \times R_s^{0.724047} \quad \text{(A-12)}$$

A.6. Petrosky Jr & Farshad (1998)

$$P_b = 112.727[(\frac{R_S^{0.5774}}{\gamma_g^{0.8439}}) \times 10^X - 12.34] \quad \text{(A-13)}$$

Where $X = 4.561 \times 10^{-5} \times T^{1.3911} - 7.916 \times 10^{-4} \times \gamma_{API}^{1.541}$

A.7. Al-Shammasi (2001)

$$P_b = \gamma_o^{5.527215} \times e^{-1.841408(\gamma_o \gamma_g)} \times [R_s \times (460+T) \times \gamma_g]^{0.783716} \quad \text{(A-14)}$$

A.8. Dindoruk & Christman (2004)

$$P_b = a_8(\frac{R_s^{a_9}}{\gamma_g^{a_{10}}} \times 10^A + a_{11}) \quad \text{(A-15)}$$

Where $A = \frac{(a_1 T^{a_2} + a_3 \gamma_{API}^{a_4})}{\left(a_5 + {}^{2R_s^{a_6}}/_{\gamma_g^{a_7}}\right)^2}$

Table A.13. Coefficients for correlation of Dindoruk & Christman (2004)

Coefficient	Value
a_1	4.86996E-6
a_2	5.730982539
a_3	9.9251E-3
a_4	1.776179364
a_5	44.2500268
a_6	2.702889206
a_7	0.744335673
a_8	3.359754970
a_9	28.10133245
a_{10}	1.579050160
a_{11}	0.928131344

A.9. Khamehchi et. al.(2009)

$$P_b = 107.93 R_s^{0.9129} \gamma_g^{-0.666} T^{0.2122} \gamma_{API}^{-1.08} \quad \text{(A-16)}$$

A.10. Arabloo et. al. (2014)

$$P_b = \frac{R_{sn}^{a_2} \times \gamma_{gn}^{a_3} \times T_n^{a_4}}{API_n} \quad \text{(A-17)}$$

Where

$R_{sn} = \frac{R_s}{(R_s+5000)}$, $T_n = \frac{T}{(T+500)}$, $API_n = \frac{\gamma_{API}}{(\gamma_{API}+50)}$ and $\gamma_{gn} = \frac{\gamma_g}{(\gamma_g+5)}$

$a_1 = 1.638783\ E10; a_2 = 0.9163610\ ; a_3 = 5.651436\ ;\ a_4 = 9.537109E-2$

A.11. Jarrahian et al. (2015)

$$P_b = P_o \gamma_o^{P_1} R_s^{P_2} T^{P_3} \gamma_g^{P_4} \times \exp(P_5 \frac{\gamma_g}{\gamma_o}) \quad \text{(A-18)}$$

Where

$P_0 = 2.99555554962301E-2; P_1 = 3.32023121093229\ ; P_2 = 8.2470515729579E-1;$
$P_3 = 1.07473639420694;\ P_4 = -4.48067373662815E-1;\ P_5 = -5.4244629566970E-1$

B. Evaluated Correlation Coefficients for Oil FVF at Bubblepoint Pressure

B.1. Standing (1947)

$$B_o = 0.9759 + 0.000120 \left[R_s \left(\frac{\gamma_g}{\gamma_o} \right)^{0.5} + 1.25T \right]^{1.2} \quad \text{(A-19)}$$

B.2. Vazquez & Beggs (1980)

For $P \leq P_b$;

$$B_o = 1 + C_1 R_s\ C_2 (T-60) \left(\frac{\gamma_o}{\gamma_{gs}} \right) +\ C_3\ R_s (T-60) \left(\frac{\gamma_o}{\gamma_{gs}} \right) \quad \text{(A-20)}$$

Where the values for the different coefficients are given as below.

Table A.14. Coefficients for correlation of Vazquez & Beggs (1980)

Coefficient	$\gamma_o \leq 30$	$\gamma_o > 30$
C_1	4.677 X 10^{-4}	4.670 x 10^{-4}
C_2	1.75 x 10^{-5}	1.100 x 10^{-5}
C_3	-1.811x 10^{-8}	1.337x10^{-9}

For $P \geq P_b$

$$B_o = B_{ob}\, exp|C_o\,(P - P_b)| \quad \text{(A-21)}$$

The oil compressibility in equation 4.0 was correlated as a function of R_s , T , γ_o , γ_g and P .

$$C_{o=}\left(a_1 + a_2\, R_s + a_3 T + a_4 \gamma_{gs} + a_5 \gamma_o\right)/a_6 P \quad \text{(A-22)}$$

Where $a_1 = -1.433.0,\ a_2 = 5.0$, $a_3 = 17.2$, $a_4 = -1180.0$, $a_5 = 12.61$ and $a_6 = 10^5$.

$$\gamma_{gs} = \gamma_{gp}\left[1 + 5.912 \times 10^{-5}\, \gamma_o\, T\, \log\frac{P}{114.7}\right]$$

Where γ_{gp} is the gas gravity obtained at separator conditions of P and T , γ_{gs} is the gas gravity(air=1) that would result from a separator conditions of 100psig.

B.3. Al-Marhoun (1988)

$$B_{ob} = 0.497069 + 0.862963 \times 10^{-3} T + 0.182594 \times 10^{-2} F + 0.318099 \times 10^{-5} F^2 \quad \text{(A-23)}$$

Where

$$F = R_s^{0.742390} \gamma_g^{0.323294} \gamma_o^{-1.202040}$$

B.4. Kartoatmodjo & Schmidt (1994)

$$B_{ob} = 0.98496 + 10^{-4} \times (R_s^{0.755} \gamma_g^{0.25} \gamma_o^{-1.5} + 0.45T)^{1.5} \quad \text{(A-24)}$$

B.5. Dokla & Osman (1992)

$$B_{ob} = 0.0431935 + 1.5667 \times 10^{-3} T + 1.39775 \times 10^{-3} M + 3.80525 \times 10^{-6} M^2 \quad \text{(A-25)}$$

Where

$$M = R_s^{0.773572} \gamma_g^{0.404020} \gamma_o^{-0.882605}$$

B.6. Al-Marhoun (1992)

$$B_{ob} = 1 + a_1 R_s + a_2 R_s\left(\gamma_g/\gamma_o\right) + a_3 R_s (T - 60) \times (1 - \gamma_o) + a_4 (T - 60) \quad \text{(A-26)}$$

Where $a_1 = 1.77342E-4$; $a_2 = 2.2016E-4$; $a_3 = 4.292580E-6$; $a_4 = 5.28707E-4$

B.7. Omar & Todd (1993)

$$B_{ob} = 0.972 + 0.000147[R_s \left(\frac{\gamma_g}{\gamma_o}\right)^{0.5} + 1.25T]^X \quad \text{(A-27)}$$

Where $X = A_1 + A_2(\gamma_{API}/\gamma_g) + A_3\gamma_g$

$A_1 = 1.1663; A_2 = 7.62E-4; A_3 = -0.0399$

B.8. Almehaideb (1997)

$$B_{ob} = 1.122018 + 1.41 \times 10^{-6} R_s T/\gamma_o^2 \quad \text{(A-28)}$$

B.9. Petrosky Jr & Farshad (1998)

$$B_{ob} = 1.0113 + 7.2046 \times 10^{-5} \times [R_s^{0.3738}(\gamma_g^{0.2914}/\gamma_o^{0.6265}) + 0.24626T^{0.5371}]^{3.0936} \quad \text{(A-29)}$$

B.10. Al-Shammasi (2001)

$$B_{ob} = 1 + 5.53 \times 10^{-7} \times R_s(T-60) + 1.81 \times 10^{-4} \times R_s/\gamma_o + B_1 \quad \text{(A-30)}$$

Where $B_1 = 4.49 \times 10^{-4} \times (T-60)/\gamma_o + 2.06 \times 10^{-4} \times (R_s\gamma_g/\gamma_o)$

B.11. Dindoruk & Christman (2004)

$$B_{ob} = a_{11} + a_{12}A + a_{13}A^2 + a_{14}(T-60)\gamma_{API}/\gamma_g \quad \text{(A-31)}$$

$$A = \frac{[\frac{R_s^{a_1}\gamma_g^{a_2}}{\gamma_0^{a_3}} + a_4(T-60)^{a_5} + a_6R_s]^{a_7}}{[a_8 + 2R_s^{a_9}(T-60)/\gamma_g^{a_{10}}]^2}$$

Table A.15. Coefficients for B_{ob} correlation of Dindoruk & Christman (2004)

Coefficient	Value		
a_1	2.510755	a_8	5.352624
a_2	-4.852538	a_9	-6.309052E-1
a_3	11.835	a_{10}	9.000749E-1
a_4	1.365428E5	a_{11}	9.871766E-1
a_5	2.25288	a_{12}	7.865146E-4
a_6	10.0719	a_{13}	2.689173E-6
a_7	4.450849E-1	a_{14}	1.100001E-5

B.12. Ikiensikimama & Ajienka (2012)

$$B_{ob} = X_1 + X_2[R_s^{X_1}\left(\frac{\gamma_g^{X_4}}{\gamma_o^{X_5}}\right) + X_6T^{X_7}]^{X_8} \quad \text{(A-32)}$$

Table A.16. Coefficients for B_{ob} correlation of Ikiensikimama & Ajienka (2012)

Coefficient	Value	Coefficient	Value
X_1	0:969581337	X_5	0:794397708
X_2	0:000116279	X_6	0:962766128
X_3	0:963558719	X_7	1:099448323
X_4	0:617166189	X_8	1:218183204

B.13. Arabloo et. al. (2014)

$$B_{ob} = 1 + a_1\left[(R_s + 2a_2)(\gamma_g + 1)log_{10}(\gamma_{API}) + a_2\right]^A \quad \text{(A-33)}$$

Where $A = a_3 + \gamma_g^{a_4}$;

$a_1 = 0.0003348062$; $a_2 = 25$; $a_3 = -0.2856905$; $a_4 = 0.03640287$

C. Evaluated Correlation Coefficients for Dead Oil Viscosity

C.1. Beggs & Robinson (1975)

$$\mu_{od} = 10^x - 1 \quad \text{(A-34)}$$

Where

$x = yT^{-1.163}$; $y = 10^z$; $z = 3.0324 - 0.02023\gamma_{API}$

C.2. Glasø (1980)

$$\mu_{od} = c(log\gamma_{API})^d \quad \text{(A-35)}$$

Where

$c = 3.141(10^{10})T^{-3.444}$; $d = 10.313(\log T) - 36.447$

C.3. Dindoruk and Christman (2004)

$$\mu_{od} = \frac{a_3 T^{a_4}(\log API)^A}{a_5 P_b^{a_6}\, a_7 R_{sb}^{a_8}} \quad \text{(A-36)}$$

Where

$A = a_1 logT + a_2$

The coefficients for the Dindoruk-Christman's dead oil viscosity correlation is given in Table 3.8.

Table A.17. Coefficients for μ_{od} correlation of Dindoruk and Christman (2004)

Coefficient	Value	Coefficient	Value
a_1	14.505357625	a_5	-3.1461171 E-09
a_2	-44.868655416	a_6	1.517652716
a_3	9.36579 E+09	a_7	0.010433654
a_4	-4.194017808	a_8	-0.000776880

C.4. Naseri et al. (2005)

$$\mu_{od} = 10^X \quad \text{(A-37)}$$

Where,

$X = 11.2699 - 4.2699 log\gamma_{API} - 2.052 logT.$

C.5. Kartoatmodjo & Schmidt (1994)

$$\mu_{od} = \left(\frac{16\times10^8}{T^{2.8177}}\right)(log\gamma_{API})^X \quad \text{(A-38)}$$

Where,

$X = 5.7536 logT - 26.9718.$

C.6. Petrosky Jr and Farshad (1998)

$$\mu_{od} = 2.3511 \times 10^7 T^{-2.10255}(log\gamma_{API})^X \quad \text{(A-39)}$$

Where,

$X = 4.59388 logT - 22.82792.$

C.7. Labedi (1992)

$$ln\mu_{od} = a_1 + a_2 ln\gamma_{API} + a_3 lnT \quad \text{(A-40)}$$

Where,

$a_1 = 21.23904, a_2 = -4.7013, a_2 = -0.6739$

C.8. Elsharkawy and Alikhan (1999)

$$\mu_{od} = 10^X - 1 \quad \text{(A-41)}$$

Where,

$X = 10^{y}$, and $y = 2.16924 - 0.02525\gamma_{API} - 0.68875logT$

D. Evaluated Correlation Coefficients for Saturated Oil Viscosity

D.1. Chew & Connally (1959)

$$\mu_{ob} = A(\mu_{od})^{b} \quad \text{(A-42)}$$

Where

$A = antilog\{R_s[2.2 \times 10^{-7}R_s - 7.4 \times 10^{-4}]\}$

$b = \frac{0.68}{10^{x}} + \frac{0.25}{10^{y}} + \frac{0.062}{10^{z}}$

And $x = 8.62 \times 10^{-5}R_s$, $= 1.1 \times 10^{-3}R_s$, $z = 3.74 \times 10^{-3}R_s$

D.2. Vazquez & Beggs (1980)

$$\mu_{ob} = X\mu_{od}^{Y} \quad \text{(A-43)}$$

Where,

$X = a_1(R_s + a_2)^{a_3}$ *;* $Y = a_4(R_s + a_5)^{a_6}$

$a_1 = 10.715$*;* $a_2 = 100$*;* $a_3 = -0.515$*;* $a_4 = 5.44$*;* $a_5 = 150$ *;* $a_6 = -0.338$.

D.3. Elsharkawy & Alikhan (1999)

$$\mu_{ob} = A(\mu_{od})^{B} \quad \text{(A-44)}$$

Where,

$A = 1241.932(R_s + 641.026)^{-1.12410}$; $B = 1768.841(R_s + 1180.335)^{-1.06622}$.

D.4. Al-Khafaji et. al. (1987)

$$\mu_{ob} = A\mu_{od}^{B} \quad \text{(A-45)}$$

$A = 0.247 + 0.2824X + 0.5657X^2 - 0.4065X^3 + 0.0631X^4$;
$B = 0.894 + 0.0546X + 0.07667X^2 - 0.0736X^3 + 0.01008X^4$;
$X = logR_s$;

D.5. Dindoruk & Christman(2004)

$$\mu_{ob} = A(\mu_{od})^{B} \quad \text{(A-46)}$$

Where

$$A = \frac{a_1}{\exp(a_2 R_s)} + \frac{a_3 R_s^{a_4}}{\exp(a_5 R_s)}$$

$$B = \frac{a_6}{\exp(a_7 R_s)} + \frac{a_8 R_s^{a_9}}{\exp(a_{10} R_s)}$$

The coefficients for the saturated oil viscosity correlation are given in Table A.18.

Table A.18. Coefficients for μ_{ob} correlation of Dindoruk and Christman (2004)

Coefficient	Value
a_1	1
a_2	4.740729 E-04
a_3	-1.023451 E-02
a_4	6.600358 E-01
a_5	1.075080 E-03
a_6	1
a_7	-2.191172 E-05
a_8	-1.660981 E-02
a_9	4.233179 E-01
a_{10}	-2.273945 E-04

D.6. Khan et. al. (1987)

$$\mu_{ob} = 0.09\gamma_g^{0.5} R_s^{\frac{1}{3}} \theta_r^{-4.5} (1-\gamma_o)^{-3} \quad \text{(A-47)}$$

Where,

$\theta_r = \frac{T+459.67}{459.67}$.

D.7. Almehaideb (1997)

$$ln\mu_{ob} = 13.4 - 0.597627 lnR_s - 0.941624 lnT - 0.555208 ln\gamma_g - 1.487449 ln\gamma_{API} \quad \text{(A-48)}$$

D.8. Labedi (1992)

$$ln\mu_{ob} = a_1 + a_2\gamma_{API} + a_3 ln\mu_{od} + a_4 lnP_b \quad \text{(A-49)}$$

Where

$a_1 = 5.397259; a_2 = -0.081557; a_3 = 0.6447; a_4 = -0.426$

E. Evaluated Correlation Coefficients for Undersaturated Oil Viscosity

E.1 Beal (1946)

$$\mu_{oa=\mu_{ob}+(P-P_b)(a_1\mu_{ob}^{a_2}+a_3\mu_{ob}^{a_4})} \quad \text{(A-50)}$$

where,

$a_1 = 24e-06; a_2 = 1.6; \ a_3 = 38e-6; a_4 = 0.56$

E.2. Vazquez & Beggs (1980)

$$\mu_0 = \mu_{ob}\left(\frac{P}{P_b}\right)^m \quad \text{(A-51)}$$

Where

$m = C_1 P^{C_2}\exp(C_3 + C_4 P)$

and $C_1 = 2.6$, $C_2 = 1.187$, $C_3 = -11.513$ and $C_4 = -8.98 \times 10^{-5}$

E.3. Labedi (1992)

$$\mu_o = \mu_{ob} + m(P - P_b) \quad \text{(A-52)}$$

Where,

$lnm = a_1 + a_2\gamma_{API} + a_3 ln\mu_{od} + a_4 lnP_b$

$a_1 = -5.728832; a_2 = -0.045361; a_3 = 0.9036; \ a_4 =$*-0.3849*

E.4. Elsharkawy and Alikhan (1999)

$$\mu_o = \mu_{ob} + 10^{-2.0771}(P - P_b)\mu_{od}^{1.19279}\mu_{ob}^{-0.40712}P_b^{-0.7941} \quad \text{(A-53)}$$

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information

UNIVERSITY OF PORTSMOUTH

Postgraduate Research Student (PGRS) Information			**Student ID:**	748008
PGRS Name:	Munirudeen Ajadi Oloso			
Department:	School of Engineering	**First Supervisor:**	Dr Mohammed Hassan	
Start Date: (or progression date for Prof Doc students)		**1 October, 2014**		

Study Mode and Route:	Part-time ☒	MPhil ☐	MD ☐
	Full-time ☐	PhD ☒	Professional Doctorate ☐

Title of Thesis:	Prediction of Reservoir Fluid Properties using Machine Learning
Thesis Word Count: (excluding ancillary data)	32424

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

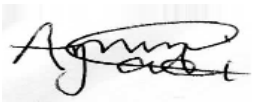
a)	Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b)	Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c)	Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d)	Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e)	Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	E840-2490-72DC-66FA-A3C9-B81A-9C29-5229

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:		**Date:** 10/12/2018